Test Design under Falsification[*]

Eduardo PEREZ-RICHET[†]   Vasiliki SKRETA [‡]

October 6, 2018

Abstract

We characterize a receiver-optimal test when manipulations are possible in the form of type falsification. Optimal design exploits the following manipulator trade-off: while falsification may lead to better grades, it devalues their meaning. We show that optimal tests can be derived among falsification-proof ones. Our optimal test has a single 'failing' grade, and a continuum of 'passing' grades. It makes the manipulator indifferent across all moderate levels of falsification. Good types never fail, but bad types may pass. An optimal test delivers at least half of the full-information value to the receiver. A three-grade optimal test also performs well.

KEYWORDS: Information Design, Falsification, Tests, Manipulation, Cheating, Persuasion.
JEL CLASSIFICATION: C72; D82.

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

1

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 persuader-receiver relationship, in which the persuader 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 type, which we assume to be either good or bad. 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 modelled 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 persuader 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, bad items generate the same signal distribution as good items. A manipulation strategy is therefore a choice of falsification rates p_B and p_G—how often, or with what probability a bad (good) item is disguised as good (bad). Good illustrations of this manipulation

---

technology are a teacher teaching a student to the test, or the way Volkswagen compromised emission tests.[3]

While this manipulation technology allows the persuader to garble the information generated by the test, and to turn any test completely uninformative, it does not make all garbles available.[4] 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 persuader 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 persuader. A rational persuader, 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. 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.[5]

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 persuader 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

---

[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]If all garbles were attainable, the persuader 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.

[5]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.

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 persuader 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 persuader 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 persuader indifferent between his optimal amount of falsification, and no falsification. The resulting test is manipulation-proof, and generates valuable information for 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 persuader 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 9

When manipulations are costly—the persuader 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 C, we show how to find an optimal test for a larger class of cost functions.

2   Related Literature

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.[6] 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 persuader'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 persuader's effort in a dynamic model where the persuader 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 persuader 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

---

[6]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.

(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) persuaders 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 makes the persuader 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.

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.

3   Model

A persuader (he) is endowed with one or multiple items. Each item is good (G) with commonly known probability μ_0 or bad (B) with probability $1 - \mu_0$ (IID in the case of multiple items). There is a test which is applied to each item. The receiver (she) decides whether to approve or
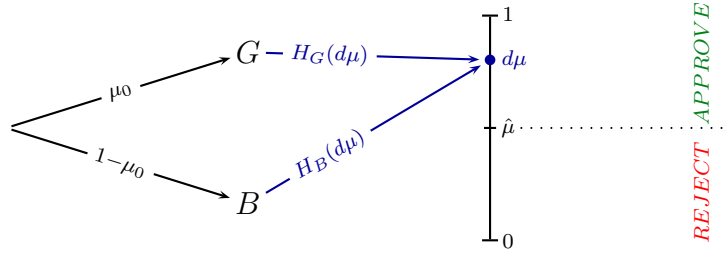
Figure 1: *A test is modelled as a Blackwell experiment. We normalize tests by equating signals to beliefs.*

reject each item after observing its test result. The persuader wants his items to be approved. His payoff from an approval is normalized to 1, and that from a rejection to 0. The receiver would like to approve only a good item: The payoff is $g > 0$ for approving a good item, and $-b < 0$ for approving a bad one. Without loss of generality, the rejection payoff is normalized to 0. Then, the receiver approves an item if she believes that it is good with probability greater than (or equal to) the threshold $\hat{\mu} = \frac{b}{g+b}$. We assume that she approves an item whenever she is indifferent.[7]

Tests. We describe the test as a Blackwell experiment (Blackwell, 1951, 1953): A measurable space of signals S, and probability measures H_G and H_B on S. A signal realization s induces a belief $\mu^s \in [0, 1]$ through Bayes' rule, where μ^s is the updated probability that the item is good. Since the approval decision only depends on the belief μ^s that the test induces, we can restrict attention to the belief distribution generated by the experiment, and denote tests by the probability measures H_G and H_B that both types generate on the space of beliefs $[0, 1]$. Then, for any measurable set $M \subseteq [0, 1]$, $H_t(M)$ is the probability that type $t \in \{G, B\}$ generates beliefs in M.

Manipulation. The persuader has access to a manipulation technology which enables type t item to generate signals according to $H_{\neg t}$ instead of H_t with some probability. The persuader chooses the probability p_t that type t items mimic type $\neg t$ for testing purposes. A manipulation

---

[7]Our analysis can be easily adapted to the case of a persuader with distinct approval values for good and bad items.

strategy is therefore a pair $(p_G, p_B) \in [0, 1]^2$. While it is natural to expect that only bad types are disguised as good types, we do not preclude good types from being disguised as bad types as part of the technology. However, we later show that it is never optimal for the persuader to do so. Figure 2 depicts the effect of manipulations on the interpretation of test-generated signals.

Timing. Given a test: First, the persuader chooses falsification rates p_G and p_B. Second, the types of the items are realized. Third, each item is subjected to the test and generates a stochastic signal s. Fourth, the receiver observes the realized signal s, forms a belief μ^s based on her knowledge of both the test and the manipulation strategy of the persuader, and finally takes an approval decision for each of them.

Thus, in the baseline model, we consider ex ante manipulations. Our analysis extends to interim manipulations by the persuader with small modifications.[8]

Solution Concept. As in Kamenica and Gentzkow (2011), our equilibrium concept is sub-game perfect equilibrium.

Falsification and Meaning of 'Grades'. With falsification, the signal μ generated by the test can no longer be equated to the belief formed by the receiver. A test (H_G, H_B) together with the persuader's falsification rates (p_G, p_B) generate a distribution of posterior beliefs of the receiver through Bayesian updating. In other words, the falsification rates and the test jointly form a new Blackwell experiment. We call this distribution of beliefs an *information structure* and denote it by F.

Modeling Assumptions. We derive the receiver-optimal test under two assumptions that allow us to focus on the main technical issues that manipulation adds to the test design problem. In Section 9, we relax both assumptions and show that the optimal test we derived is still optimal if the persuader has a continuum of IID items or if receivers approve items sequentially.

Assumption 1 (Perfect Observability). *The falsification rates p_B and p_G are observed by the receiver before she makes her approval decisions.*

---

[8]Details of this analysis are available from the authors upon request.

Assumption 2 (Falsification Rates Bound)**.** *The persuader is restricted to falsification rates such that $p_B + p_G \leq 1$.*

Without Assumption 1 correct inference occurs only on the equilibrium path but with this assumption, beliefs are correctly updated beliefs off-path as well. Note that Assumption 2 is satisfied in particular when the persuader can only, or is only willing to, disguise bad types as good types, so $p_G = 0$. When Assumption 2 holds, higher signals correspond to higher true beliefs. If the persuader could choose falsification rates that do not satisfy Assumption 2, this would lead to a reversal of the meaning of signals as higher signals would lead to lower beliefs. This assumption is important under Assumption 1, as the optimal test we derive in the first part of the paper under Assumption 1 and Assumption 2 will not be immune to deviations such that $p_B + p_G > 1$ (see Appendix B). However, it is irrelevant in the interpretation of the model where we relax Assumption 1 in Section 9, as imperfect observability ensures that such deviations can be discouraged. We elaborate on this in Section 9.

Next, we make several comments about the model that help clarify the role of these assumptions, and the consequences of our modeling choices.

Discussion of the Model. First, we discuss the manipulation technology. Note that 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 persuader to render *any* test uninformative by choosing $p_B + p_G = 1$. If $\mu_0 \geq \hat{\mu}$, so that the receiver approves when her belief is equal to the prior, making the test uninformative is actually the optimal choice of the persuader. This is why, in what follows, we focus on the interesting case where $\mu_0 < \hat{\mu}$. For a given test, however, the persuader 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 persuader could generate any such garbling, then the optimal design problem would always result in the optimal information structure of the persuader. Then, we can view the problem of the persuader in our setup as "constrained Bayesian persuasion:" the test and the falsification technology together induce a constrained set of information structures among which the persuader can choose freely.

The reason we picked this technology is because it is natural and fits well a number of examples mentioned in the introduction. However, other choices might be interesting as well.

Presumably, any choice of manipulation technology would specify the ways in which tests can be garbled and the cost of doing so. If no restrictions were put on available garbles, the optimal test design problem would be moot as it would always result in the persuader-optimal information structure, that is the solution of the Bayesian persuasion problem (Kamenica and Gentzkow, 2011) where the persuader is the sender. This is because any test that is more informative than the sender-optimal one would be garbled back to it, whereas any other test would result in an even worse information structure for the receiver.

Because too much falsification leads the receiver to beliefs that punish the persuader by lowering approval rates, costs are not needed to create a trade-off for the persuader that test design can exploit. Studying the problem without costs allows us to understand the effect of this trade-off more purely. Interestingly, we find that the absence of costs does not lead the persuader to make the test completely uninformative when $\mu_0 < \hat{\mu}$. However, a natural extension of our falsification technology is to make it costly. Indeed, costs can capture inherent technological costs, as well as expected fines that a manipulator may have to pay if caught, and/or ethical and emotional discomfort. We study costly falsification in Section 8.

Next, we comment on the lack of commitment assumption by receivers in our baseline model. With commitment and 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 introduced in Section 8. Suppose, for example, that the regulator is willing to punish the persuader 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. Section 8 shows what can be achieved with lower expected punishments, and derives a lower bound on costs for

full information to be achievable.

Finally, we discuss the perfect observability assumption. It is a simplifying assumption that captures the idea that, receivers often have a good understanding of the amount of manipulation they are facing. Interestingly, in equilibrium it is also in the persuader's interest to commit to observable manipulations even if such manipulations are perceived as bad. The persuader benefits from observability in the same way the sender benefits from commitment in the usual Bayesian persuasion case. To see this, consider the case where the falsification rates are not observable. Then our problem can be formulated as a mediation problem,[9] where the receiver-optimal design problem is that of a mediator taking reports from the persuader, and making recommendations to the receiver. In this case, it is easy to see that the mediator cannot generate any information. Indeed, to make truthful reporting by the persuader incentive compatible, she must recommend approval with the same probability for good and bad items, therefore she cannot convey any information to the receiver, and her recommendation must be to always reject since $\mu_0 < \hat{\mu}$. But then, this means that the persuader can only benefit from observability.

Assumption 1 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 Section 9. 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.

4   Examples and Benchmarking

Binary Tests.   The receiver would like to be perfectly informed about the types of items. But if the test is fully informative, the persuader has an incentive to falsify. In fact, faced with

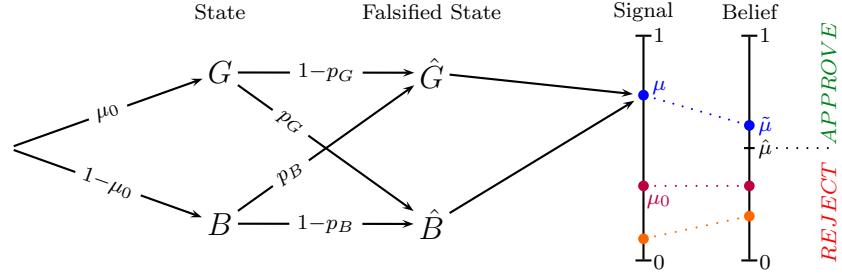---

[9]See Myerson (1991, Chapter 6).

Figure 2: *The effect of falsification on beliefs under Assumption 1 and Assumption 2.*

a fully informative test, the persuader finds herself in the shoes of the sender in the Bayesian persuasion model of Kamenica and Gentzkow (2011). He chooses $p_G = 0$ and $p_B = \frac{\mu_0(1-\hat{\mu})}{\hat{\mu}(1-\mu_0)}$, so that, when the receiver sees signal $\mu = 1$, the belief she forms is exactly equal to $\hat{\mu}$. We refer to the resulting information structure as the KG information structure, and to the associated payoffs as the KG payoffs. The persuader's KG payoff is $\mu_0 + (1 - \mu_0)p_B = \frac{\mu_0}{\hat{\mu}}$, which is the highest possible payoff she can obtain, whereas the receiver's KG payoff is 0, as in the absence of information.

In many information acquisition/transmission frameworks in which the action is binary, a revelation-principle result holds which says that one can, without loss of generality, restrict attention to binary experiments. This is not the case here, but it is interesting to consider what happens with binary signals. Whenever a binary test is more informative than the KG information structure, the persuader falsifies so as to garble it into the KG information structure. Indeed, such a test generates two signals: A low signal $\mu = 0$, and a high signal $\overline{\mu}$ above the threshold $\hat{\mu}$, where a good type generates the high signal $\overline{\mu}$ with probability 1, and a bad type generates $\overline{\mu}$ with probability $\pi_B < \mu_0 \frac{1-\hat{\mu}}{1-\mu_0}$. But then the persuader obtains the KG payoff by choosing p_B so as to make the probability that a bad type generates the high signal $p_B + (1 - p_B)\pi_B$ equal to $\mu_0 \frac{1-\hat{\mu}}{1-\mu_0}$, that is $p_B = \frac{1}{1-\pi_B}\left(\mu_0 \frac{1-\hat{\mu}}{1-\mu_0} - \pi_B\right)$. Hence, the receiver gets a payoff of 0. If, instead, a binary test is less informative than, or not comparable with the KG information structure, the payoff of the persuader is below his KG payoff, but the receiver payoff is not increased. Thus, we have proved the following result.

Proposition 1 (Binary Tests)**.** *With binary tests, the receiver always gets a payoff of 0. If the test is more informative than the KG information structure, the persuader gets his KG payoff. Otherwise, the payoff of the persuader is strictly below his KG payoff.*
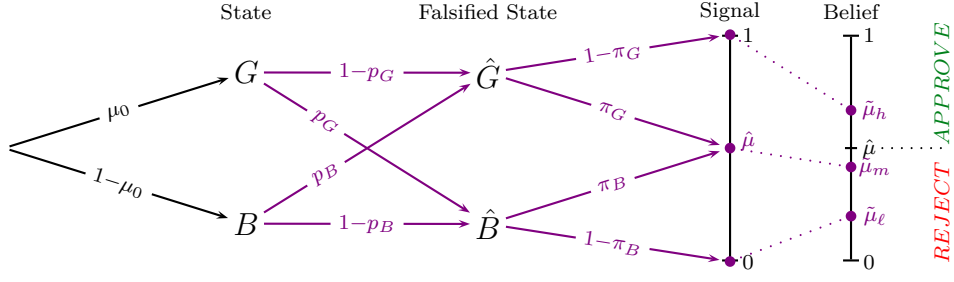
Figure 3: *A Better Test. The signal column corresponds to beliefs in the absence of falsification, the belief column gives the belief associated with each signal when there is falsification.*

A Better Test. Consider the test described in Figure 3, and recall that signals correspond to beliefs in the absence of falsification. This test has high signal generated only by G, so this signal is equal to 1, a low signal only generated by B, so it is equal to 0, and a middle signal equal to $\hat{\mu}$ generated by both G and B, with respective probabilities π_G and π_B. We pick $\pi_G = \frac{(1-\mu_0)\hat{\mu}}{\mu_0(1-\hat{\mu})}\pi_B > \pi_B$, so that the belief corresponding to the middle signal in the absence of falsification is indeed equal to $\hat{\mu}$. When the persuader falsifies, the receiver associates new beliefs to each of the three signals. These beliefs are

$$\tilde{\mu}_h = \frac{\mu_0(1 - p_G)}{\mu_0(1 - p_G) + (1 - \mu_0)p_B},$$

$$\tilde{\mu}_m = \frac{\mu_0\pi_G - \mu_0(\pi_G - \pi_B)p_G}{\mu_0\pi_G + (1 - \mu_0)\pi_B - \mu_0(\pi_G - \pi_B)p_G + (1 - \mu_0)(\pi_G - \pi_B)p_B},$$

$$\tilde{\mu}_\ell = \frac{\mu_0 p_G}{\mu_0 p_G + (1 - \mu_0)(1 - p_B)}.$$

Simple calculations show that $\tilde{\mu}_h$, and, more importantly, $\tilde{\mu}_m$, are decreasing in both p_G and p_B, whereas $\tilde{\mu}_\ell$ is increasing in both. Therefore any small amount of falsification implies that an item is no longer approved when the receiver receives the middle signal $\hat{\mu}$, as the corresponding belief falls below $\hat{\mu}$. The only benefit from falsification is therefore to increase the probability that a bad type generates the high signal by increasing p_B. Increasing p_G, however, is only harmful, so the persuader sets $p_G = 0$. The maximum and optimal level of p_B is the one that brings $\tilde{\mu}_h$ down to $\hat{\mu}$, since falsifying more than this would lead the receiver to approve none of the items. Let $\overline{p}_B = \frac{\mu_0(1-\hat{\mu})}{(1-\mu_0)\hat{\mu}}$ denote this level. The payoff of the persuader if he chooses this

13

maximum falsification level $\overline{p}_B$ is

$$\big(\mu_0 + (1 - \mu_0)\overline{p}_B\big)(1 - \pi_G) = \frac{\mu_0}{\hat{\mu}} - \frac{1 - \mu_0}{1 - \hat{\mu}}\pi_B,$$

while her no-falsification payoff is

$$\mu_0 + (1 - \mu_0)\pi_B.$$

The test can discourage falsification by equating the two, which is achieved by choosing $\pi_B^* = \frac{\mu_0(1-\hat{\mu})^2}{(1-\mu_0)\hat{\mu}(2-\hat{\mu})}$, and $\pi_G^* = \frac{1-\hat{\mu}}{2-\hat{\mu}}$. This test gives the receiver a payoff of

$$\mu_0 g - (1 - \mu_0)\pi_B^* b = (g + b)\frac{\mu_0(1 - \hat{\mu})}{2 - \hat{\mu}} > 0.$$

These observations are summarized in the following:

Proposition 2. *The test described in Figure 3 with π_B^* and π_G^* gives the persuader no incentive to falsify, and yields a strictly positive payoff for the receiver.*

Intuitively, enriching the set of signals by adding a middle signal $\hat{\mu}$ makes the persuader unwilling to falsify, as any falsification would lead the receiver to devalue the middle signal, and no longer approve items that generate this signal. This test, while not perfectly informative, enables the generation of useful information despite the possibility of costless falsification. Hence, the curse of falsification can be beaten by good design.

We can think of several testing procedures that would generate this information structure. One is to use a perfectly informative test, and simply garble the results provided to the receiver. Another possibility is to design two pass-fail tests to which items would be randomly and independently assigned: the first pass-fail test, assigned with probability $1 - \pi_G^*$, is perfectly informative about the type, and the other one, assigned with probability π_G^*, is such that the good type passes with probability one, and the bad type with probability π_B^*/π_G^*, so that a pass in this state leads to belief $\hat{\mu}$. In this implementation, manipulations lead the receiver to reject all items subjected to the second test, regardless of the outcome. In the remainder of the paper, we proceed to find a receiver-optimal test.

14

5 Tests and Information Structures

To proceed with the general analysis, we employ a useful representation of experiments as convex functions that, to our knowledge, first appears in Kolotilin (2016), and is also discussed at length in Gentzkow and Kamenica (2016b).

Bayesian Consistency. We denote by F both a probability measure on $[0, 1]$ and the corresponding pseudo cdf,[10] so $F(\mu)$ and $F\big([0, \mu)\big)$ are used interchangeably. It is a posterior belief distribution if and only if $\int_0^1 \mu F(d\mu) = \mu_0$ (see Kamenica and Gentzkow, 2011) or, equivalently, integrating by parts,

$$\int_0^1 F(\mu) d\mu = 1 - \mu_0. \tag{BC}$$

Experiments as Convex Functions. For a belief distribution F that satisfies (BC), we can define the function

$$\mathcal{F}(\mu) = \int_0^\mu F(x) dx$$

from $[0, 1]$ to $[0, 1 - \mu_0]$. Let Δ^B be the set of increasing convex functions of μ on $[0, 1]$ that are bounded above by $(1 - \mu_0)\mu$, and below by $(\mu - \mu_0)^+$. This set is illustrated in Figure 4. Then $\mathcal{F}(\cdot) \in \Delta^B$. Reciprocally, any function $\mathcal{F} \in \Delta^B$ admits a left derivative that is the pseudo cdf of a Bayes consistent belief distribution. Therefore, there is a one-to-one relationship between functions in Δ^B and Bayes consistent belief distributions. The upper bound on Δ^B corresponds to the pseudo cdf $F(\mu) = 1$, which is the fully informative experiment. The lower bound on Δ^B corresponds to the pseudo cdf $F(\mu) = \mathbb{1}_{\mu > \mu_0}$, which corresponds to the uninformative experiment and puts probability one on the prior μ_0. The following lemma states this characterization, and is proved in Appendix A.

Lemma 1. $\mathcal{F} \in \Delta^B$ if and only if there exists a Bayes consistent belief distribution F such that, for all $\mu \in [0, 1]$, $\mathcal{F}(\mu) = \int_0^\mu F(x) dx$.

---

[10]If F is a probability measure on the space of beliefs $[0, 1]$, then it has a cumulative distribution function $\tilde{F} : [0, 1] \to [0, 1]$. Slightly abusing notations, we then denote the pseudo cdf of a probability measure F by the same letter F, and define it for $\mu \in (0, 1]$ by $F(\mu) = \sup_{x < \mu} \tilde{F}(x)$. Hence, for $\mu > 0$, $F(\mu)$ is the probability measure of the set $[0, \mu)$. For example, in a perfectly informative information structure, a good item generates belief 1 with probability 1, and the bad type generates belief 0 with probability 1, that is $F_G(\mu) = 0$ and $F_B(\mu) = 1$ for all $\mu \in (0, 1]$. In a perfectly uninformative experiment, both types generate belief μ_0 with probability 1, that is $F_G(\mu) = F_B(\mu) = \mathbb{1}_{\mu > \mu_0}$.
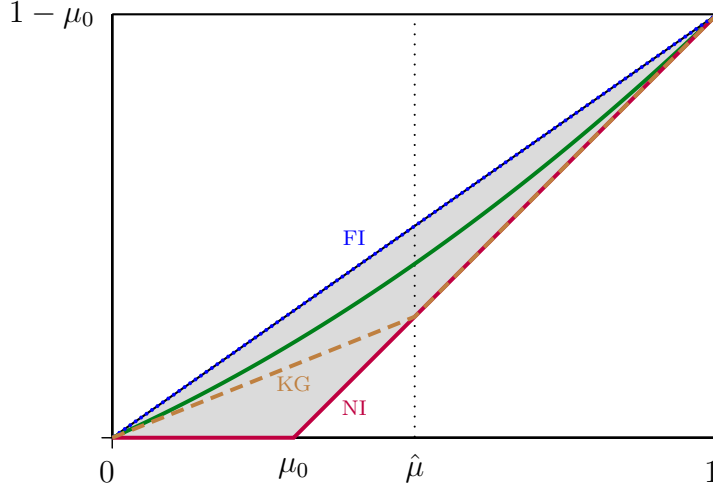
Figure 4: Δ^B *is the set of increasing convex functions in the grey triangle– the green curve is an example of a function in* Δ^B, *the brown dashed kinked line corresponds to the KG information structure which obtains when the test is fully informative, the top dotted blue line corresponds to full information (FI), the bottom kinked line corresponds to no information (NI). In this and all subsequent figures, we take* $\mu_0 = 0.3$ *and* $\hat{\mu} = 0.5$.

We can re-express the distributions of beliefs induced by good and bad types as functions of the posterior belief distribution F.

Lemma 2. *The belief distributions generated by the good type and the bad type are respectively*

$$F_G(\mu) = \frac{1}{\mu_0}\Big\{\mu F(\mu) - \mathcal{F}(\mu)\Big\},$$

$$F_B(\mu) = \frac{1}{1-\mu_0}\Big\{(1-\mu)F(\mu) + \mathcal{F}(\mu)\Big\}.$$

In the absence of falsification a test H induces an information structure, and thus satisfies Lemma 2 with the representation $\mathcal{H}$. In the presence of falsification, the test H still satisfies these relationships, that is, we have, for each signal $\mu \in (0,1]$,

$$H_G(\mu) = \frac{1}{\mu_0}\Big\{\mu H(\mu) - \mathcal{H}(\mu)\Big\},$$

and

$$H_B(\mu) = \frac{1}{1-\mu_0}\Big\{(1-\mu)H(\mu) + \mathcal{H}(\mu)\Big\}.$$

However, as already explained, the signals generated by $\mathcal{H}$ are no longer beliefs when there is

16

falsification.

Modified Payoffs. We can obtain convenient expressions of the players' payoffs using $\mathcal{F}$. The payoff of the persuader is given by the probability that he generates a belief above the threshold, $1 - F(\hat{\mu})$. Graphically, the persuader would like the left derivative $F(\hat{\mu})$ of $\mathcal{F}$ at $\hat{\mu}$ to be as small as possible. The payoff of the receiver, scaled by $\frac{1}{g+b}$, is

$$\frac{1}{g+b} \int_{\hat{\mu}}^{1} \big(\mu g + (1-\mu)(-b)\big) F(d\mu) = 1 - \hat{\mu} - \int_{\hat{\mu}}^{1} F(x)dx$$

$$= \mu_0 - \hat{\mu} + \mathcal{F}(\hat{\mu}).$$

Since the constant terms are irrelevant for optimization, we use $\mathcal{F}(\hat{\mu})$ as our objective function. This objective function is easily pictured in Figure 4, and it appears clearly that, in the absence of any falsification constraints, the receiver-optimal information structure would be the upper-bound function of Δ^B, which corresponds to full information (FI). It is easy to see on Figure 4 why the KG information structure is optimal for the persuader, and pessimal for the receiver, whereas full information is optimal for the receiver. No information (NI) is pessimal for both. The payoff space generated by all possible information structures is illustrated on Figure 11, below.

6 Optimal Approval and Optimal Falsification

Optimal Approval. To understand the incentives of the persuader to falsify, we start by describing how falsification affects the receiver's approval decisions. If the persuader decides to falsify, he changes the belief associated with each signal. Let μ be both the signal received by the receiver, and the belief she forms in the absence of falsification. Then, if the persuader chooses a falsification strategy (p_B, p_G), the receiver forms belief $\tilde{\mu} \neq \mu$ when she receives signal μ. Their relationship, which we call the *belief transformation*, is stated in closed form in the next lemma and holds for *all* values of p_B and p_G, that is, even without the restriction of Assumption 2. Interestingly, the belief transformation is independent of the test, and depends only on the falsification strategy. Hence, any falsification strategy induces a reinterpretation of signals that does not depend on the test.
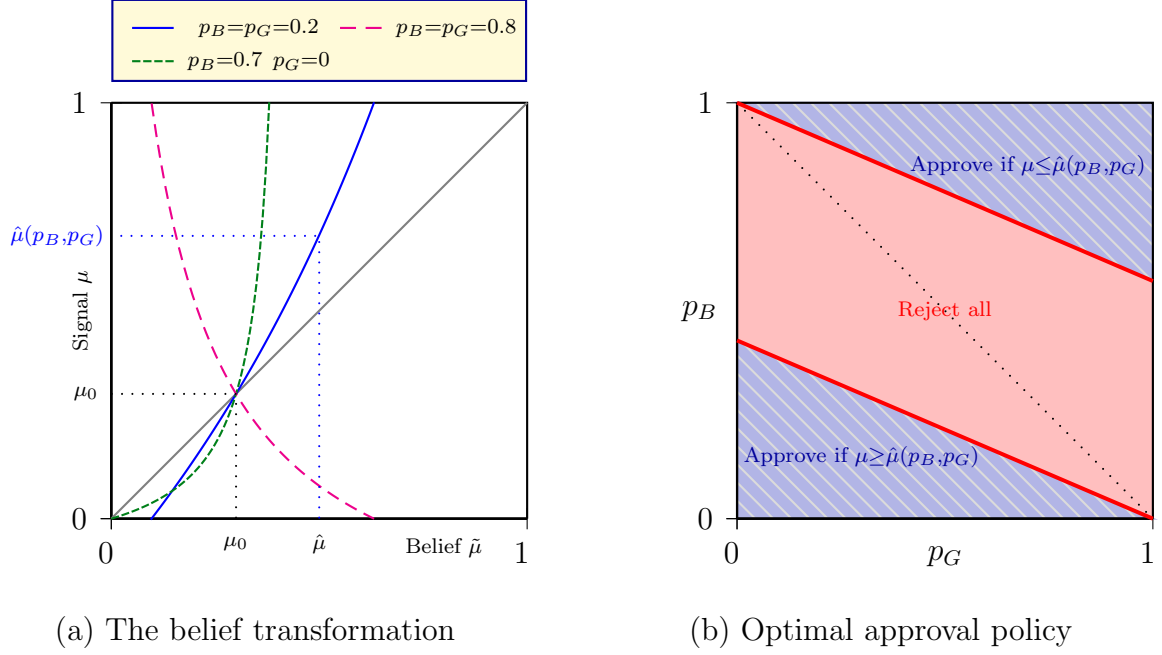
(a) The belief transformation

(b) Optimal approval policy

Figure 5: *Panel (a) illustrates the relationship between signal (or pre-falsification belief), and actual (post-falsification) belief. Panel (b) illustrates the optimal approval policy: the red line is the line with equation $p_B = \frac{\mu_0(1-\hat{\mu})}{\hat{\mu}(1-\mu_0)}(1-p_G)$; in the solid pink region above the red line, the receiver never approves; in the hatched blue region below the red line, she uses an approval threshold $\hat{\mu}(p_B, p_G)$.*

Lemma 3 (Belief Transformation). *Under Assumption 1, with falsification (p_B, p_G), signal μ induces belief $\tilde{\mu}$, where*

$$\mu = \mu_0 \frac{(1-\mu_0)\tilde{\mu} - \mu_0(1-\tilde{\mu})p_G - (1-\mu_0)\tilde{\mu}p_B}{\mu_0(1-\mu_0) - \mu_0(1-\tilde{\mu})p_G - (1-\mu_0)\tilde{\mu}p_B}. \tag{BT}$$

This function has a fixed point μ_0. It is increasing in $\tilde{\mu}$ if $p_B + p_G < 1$, decreasing if $p_B + p_G > 1$, and constant to μ_0 otherwise. The range of beliefs $\tilde{\mu}$ is the interval $\left[\underline{\mu}, \overline{\mu}\right]$, where $\underline{\mu} = \frac{\mu_0 p_G}{\mu_0 p_G + (1-\mu_0)(1-p_B)}$, and $\overline{\mu} = \frac{\mu_0(1-p_G)}{\mu_0(1-p_G) + (1-\mu_0)p_B}$.

If the amount of falsification is constrained by Assumption 2, the receiver still associates higher signals μ with higher beliefs $\tilde{\mu}$, but this is reversed when $p_B + p_G > 1$. The belief transformation is illustrated in panel (a) of Figure 5 for different values of p_B and p_G. Note that, with falsification, beliefs may be bounded away from 0 or 1. Whenever $p_B > 0$, the receiver can never be sure that she is facing a bad type, and whenever $p_G > 0$, she can never be sure that she is facing a good type.

The receiver approves when her belief exceeds $\hat{\mu}$, that is when her signal μ exceeds the

threshold $\hat{\mu}(p_B, p_G)$ obtained from the belief transformation, as illustrated by the first curve of panel (a) in Figure 5. For some values of (p_B, p_G), such signals cannot be generated (this is the case when $\overline{\mu} < \hat{\mu}$), and the receiver never approves, as illustrated by the second curve of panel (a) in Figure 5. The following proposition characterizes the optimal approval strategy under falsification.

Proposition 3 (Optimal Approval). *Under Assumption 1, there exists a threshold*

$$\hat{\mu}(p_B, p_G) = \mu_0 \frac{(1 - \mu_0)\hat{\mu} - \mu_0(1 - \hat{\mu})p_G - (1 - \mu_0)\hat{\mu}p_B}{\mu_0(1 - \mu_0) - \mu_0(1 - \hat{\mu})p_G - (1 - \mu_0)\hat{\mu}p_B},$$

such that:

(i) *If $p_B < \frac{\mu_0(1-\hat{\mu})}{\hat{\mu}(1-\mu_0)}(1 - p_G)$, $\hat{\mu}(p_B, p_G)$ is increasing in p_B and p_G, and the receiver approves any item generating a signal $\mu \geq \hat{\mu}(p_B, p_G)$.*

(ii) *If $p_B > 1 - \frac{\mu_0(1-\hat{\mu})}{\hat{\mu}(1-\mu_0)}p_G$, $\hat{\mu}(p_B, p_G)$ is decreasing in p_B and p_G, and the receiver approves any item generating a signal $\mu \leq \hat{\mu}(p_B, p_G)$.*

(iii) *Otherwise, the receiver rejects every item.*

The optimal policy is illustrated in panel (b) of Figure 5. Note that $\hat{\mu}(0, 0) = \hat{\mu}$ as, then, signals coincide with beliefs.

Optimal Falsification. Now, consider the problem of the persuader under both Assumption 1 and Assumption 2. Whenever there is falsification, the threshold $\hat{\mu}(p_B, p_G)$ is higher than $\hat{\mu}$. Since the threshold is increasing in p_B and p_G, more falsification hurts both types as it makes the receiver more selective. However, it also changes the probabilities with which both types generate the different signals in a way that can benefit the persuader. To see this, we compute the persuader's falsification payoff. It is 0 in the region where the receiver rejects for all signals. In the threshold region, we can write the persuader's payoff as

$$\Pi(p_B, p_G) = 1 - \big\{\mu_0(1-p_G) + (1-\mu_0)p_B\big\} H_G\big(\hat{\mu}(p_B, p_G)\big) - \big\{\mu_0 p_G + (1-\mu_0)(1-p_B)\big\} H_B\big(\hat{\mu}(p_B, p_G)\big).$$
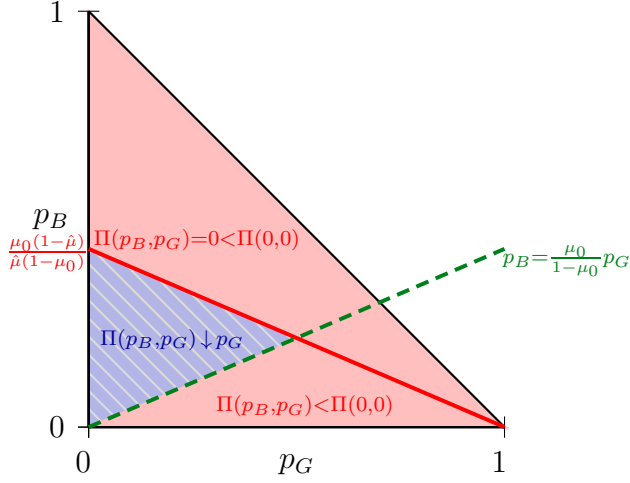
19

Figure 6: *Optimal falsification under Assumption 1 and Assumption 2 if $H(\hat{\mu}) < 1$.*

Using the expressions from Lemma 2 applied to H_G and H_B, we obtain

$$\Pi(p_B, p_G) = 1 - H\big(\hat{\mu}(p_B, p_G)\big) + \left(\frac{p_B}{\mu_0} - \frac{p_G}{1 - \mu_0}\right)\left\{\mathcal{H}\big(\hat{\mu}(p_B, p_G)\big) - \big(\hat{\mu}(p_B, p_G) - \mu_0\big)H\big(\hat{\mu}(p_B, p_G)\big)\right\}. \tag{1}$$

This expression, as we show, implies that, in any optimal falsification strategy that follows a relevant test, $p_G = 0$. Intuitively, pretending that items are bad when in fact they are good not only increases the approval threshold, but also deteriorates the signal distribution generated by good types. It may, however, be payoff-improving for the persuader to sometimes pretend that an item is good when in fact it is bad. Even though it increases the approval threshold, it allows bad items to generate the same signal distribution as good ones, and therefore be approved with a higher probability.

Proposition 4 (Optimal Falsification). *Under Assumption 1, and Assumption 2, any optimal falsification strategy satisfies the following.*

   (i) If $H(\hat{\mu}) < 1$, then $p_G = 0$ and $p_B \leq \frac{\mu_0(1-\hat{\mu})}{\hat{\mu}(1-\mu_0)}$.

   (ii) If $H(\hat{\mu}) = 1$, then falsification is inconsequential, the receiver never approves, and all players get a null payoff.

The idea of the proof, can be visualized on Figure 6. First, we show that all falsification

strategies that do not lie in the hatched triangle are dominated by no falsification. Second, we show that $\Pi(p_B, p_G)$ is decreasing in p_G within the hatched triangle.

Proposition 4 implies that the optimal falsification problem can be reduced to the choice of $p_B \in \left[0, \frac{\mu_0(1-\hat{\mu})}{\hat{\mu}(1-\mu_0)}\right]$, thus generating an approval threshold $\hat{\mu}(p_B, 0)$ between $\hat{\mu}$ and 1. We can reformulate this problem as the choice of a threshold $\mu \in [\hat{\mu}, 1]$, and invert the function $\hat{\mu}(p_B, 0)$ to get the level of falsification p_B that corresponds to a threshold μ,

$$p_B = \frac{\mu_0(\mu - \hat{\mu})}{\hat{\mu}(\mu - \mu_0)}.$$

Replacing this in (1), we obtain the falsification payoff of the persuader as a function of the induced signal threshold μ

$$
\begin{aligned}
\Pi(\mu) &= 1 - H(\mu) + \frac{\mu - \hat{\mu}}{\hat{\mu}(\mu - \mu_0)}\Big\{\mathcal{H}(\mu) - (\mu - \mu_0)H(\mu)\Big\} \\
&= 1 + \frac{\mu - \hat{\mu}}{\hat{\mu}(\mu - \mu_0)}\mathcal{H}(\mu) - \frac{\mu}{\hat{\mu}}H(\mu),
\end{aligned}
\tag{2}
$$

and the optimal falsification problem reduces to choosing which approval threshold to induce so as to maximize $\Pi(\mu)$ on $[\hat{\mu}, 1]$.

7 Optimal Design

We now consider the problem of designing a receiver-optimal test in the presence of falsification, under Assumption 1 and Assumption 2. Both these assumptions are relaxed in Section 9.

A No-Falsification Principle. We start by showing that a no-falsification principle holds. It states that any final information structure and, therefore, any payoffs that can be generated with falsification, can also be generated without falsification. The logic of the argument is similar to that of the revelation principle. Consider any test, and the optimal falsification strategy of the persuader associated with this test. Together, they generate a certain information structure. Now, consider the test that generates this precise information structure, instead of the initial test. Then, the persuader has no incentive to falsify under this new test. The main difference with the usual revelation principle is in the link between deviations from no falsification under
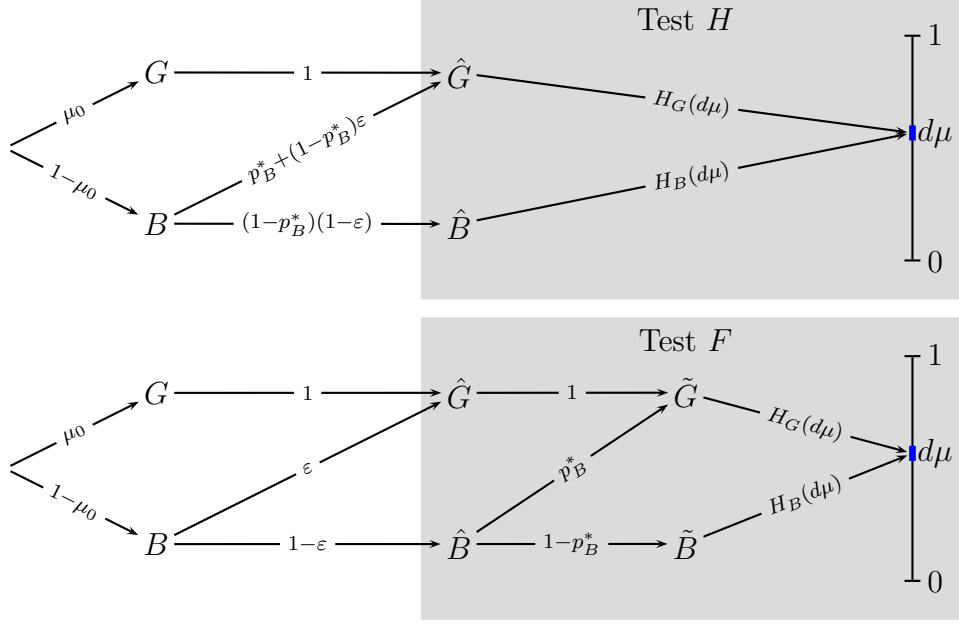
Figure 7: *Experiment and final information structure with p_B^*.*

the new test, and corresponding deviations from the optimal level of falsification under the initial test.

More formally, consider a test H, and let $p_B^* > 0$ be the associated optimal falsification strategy of the persuader. Together, p_B^* and H define a new experiment, characterized by a posterior belief distribution F. One way to deliver this experiment, is to choose the test F described in the lower panel of Figure 7. As illustrated by Figure 7, falsifying by choosing $p_B = \varepsilon$ under this new test induces the same posterior belief distribution as increasing the level of falsification by $(1 - p_B^*)\varepsilon$ under the initial test H. But since p_B^* is optimal under H, this deviation must be unprofitable to the persuader. Therefore, it is optimal for the persuader not to falsify under the new test F. This proves the no-falsification principle,[11] which we now state more formally.

Proposition 5 (No-Falsification Principle)**.** *If a test can induce a final belief distribution F with falsification $p_B^* > 0$, then this distribution can also be induced by a falsification-proof test. In both cases, the receiver payoff is given by $\mathcal{F}(\hat{\mu})$, and the persuader payoff by $1 - F(\hat{\mu})$.*

---

[11]The no-falsification principle holds for any state space (not just binary as in our model) so long as falsification is costless or falsification costs are concave in falsification rates. Details are available from the authors upon request.

Optimal Design.   The no-falsification principle implies that we can restrict the optimal design problem to the one of finding an optimal test under which the persuader has no incentive to falsify. A test H is such that the persuader has no incentive to falsify if and only if $\Pi(\hat{\mu}) \geq \Pi(\mu)$, for all $\mu \in [\hat{\mu}, 1]$, that is, recalling the payoff formula (2), if and only if H satisfies the following incentive constraint

$$\frac{\mu - \hat{\mu}}{\mu - \mu_0}\mathcal{H}(\mu) \leq \mu H(\mu) - \hat{\mu}H(\hat{\mu}), \qquad \forall \mu \in \big[\hat{\mu}, 1\big]. \tag{IC$_0$}$$

And, if this is the case, the payoff of the receiver is given by $\mathcal{H}(\hat{\mu})$ (up to constants). Hence the receiver-optimal design problem is

$$\max_{\mathcal{H} \in \Delta^B} \mathcal{H}(\hat{\mu})$$
$$\text{s.t. } \frac{\mu - \hat{\mu}}{\mu - \mu_0}\mathcal{H}(\mu) \leq \mu H(\mu) - \hat{\mu}H(\hat{\mu}), \qquad \forall \mu \in \big[\hat{\mu}, 1\big]. \tag{IC$_0$}$$

To form intuition about this program, it is useful to go back to Figure 4. We want to maximize $\mathcal{H}(\hat{\mu})$ subject to a constraint on the values taken by $\mathcal{H}$ to the right of $\hat{\mu}$. There is no incentive constraint on $\mathcal{H}$ to the left of $\hat{\mu}$. Recall that $H(\hat{\mu})$ is the left-derivative of $\mathcal{H}$ at $\hat{\mu}$.

A first remark is that we can look for optimal tests that are linear to the left of $\hat{\mu}$. To see this, suppose that $\mathcal{H} \in \Delta^B$ satisfies (IC$_0$), and consider the function

$$\tilde{\mathcal{H}}(\mu) = \begin{cases} \mu\mathcal{H}(\hat{\mu})/\hat{\mu} & \text{if } \mu \leq \hat{\mu} \\ \mathcal{H}(\mu) & \text{if } \mu \geq \hat{\mu} \end{cases}.$$

It is easy to see that $\tilde{\mathcal{H}}$ is in Δ^B, and since $\tilde{H}(\hat{\mu}) = \mathcal{H}(\hat{\mu})/\hat{\mu} \leq H(\hat{\mu})$, by convexity of $\mathcal{H}$, the new experiment $\tilde{\mathcal{H}}$ also satisfies (IC$_0$), and delivers the same payoff to the receiver. Therefore, we have proved the following lemma.

Lemma 4. *For every test $\mathcal{H}$ that satisfies* (IC$_0$), *there is a test $\tilde{\mathcal{H}}$ that is linear to the left of $\hat{\mu}$, satisfies* (IC$_0$), *and delivers the same payoff to the receiver.*

Linearity means that we can look for optimal tests that put an atom on belief 0, and never generate any belief in $(0, \hat{\mu})$. In particular, we can restrict ourselves to tests such that good types are never rejected. Another consequence of Lemma 4 is that we can look for optimal

23

tests that are on the Pareto frontier. Indeed, recalling the definition of the set Δ^B, it is easy to visualize on Figure 4 that $\tilde{\mathcal{H}}$ is the test with the lowest possible left derivative at $\hat{\mu}$ among tests that deliver payoff $\mathcal{H}(\hat{\mu})$ to the receiver.

Next, we denote the left derivative of $\mathcal{H}$ at $\hat{\mu}$ by κ. Since $\mathcal{H} \in \Delta^B$, we must have $0 \leq \kappa \leq 1 - \mu_0$. Note that the (IC$_0$) constraint is automatically satisfied at $\hat{\mu}$. Therefore, we can rewrite it as

$$\mu H(\mu) - \frac{\mu - \hat{\mu}}{\mu - \mu_0} \mathcal{H}(\mu) \geq \kappa\hat{\mu}, \qquad \forall \mu > \hat{\mu}. \tag{IC$_0'$}$$

Then, the optimal design problem reduces to choosing $\kappa \in [0, 1 - \mu_0]$, and $\mathcal{H} \in \Delta^B$ such that $\mathcal{H}(\mu) = \kappa\mu$ for $\mu \leq \hat{\mu}$ so as to maximize κ, under the constraint (IC$_0'$).

As a first exercise, we can find the receiver-optimal test with three signals, and compare it to the test we described in Section 4. This test must be linear to the right of $\hat{\mu}$. Let η be its slope to the right of $\hat{\mu}$. We must have $\eta = \frac{1 - \mu_0 - \kappa\hat{\mu}}{1 - \hat{\mu}}$. And we can rewrite (IC$_0'$) as

$$\eta\mu - \frac{\mu - \hat{\mu}}{\mu - \mu_0} \big(\kappa\hat{\mu} + \eta(\mu - \hat{\mu})\big) \geq \kappa\hat{\mu}, \qquad \forall \mu > \hat{\mu}.$$

A quick calculation shows that the left-hand side is strictly decreasing in μ. So the incentive constraint can be simplified to

$$\eta - \frac{1 - \hat{\mu}}{1 - \mu_0} \big(\kappa\hat{\mu} + \eta(1 - \hat{\mu})\big) \geq \kappa\hat{\mu}.$$

Replacing η by its expression, and rearranging, we obtain

$$\kappa \leq \frac{(1 - \mu_0) - (1 - \hat{\mu})^2}{\hat{\mu}(2 - \hat{\mu})}.$$

Since we want to maximize $\mathcal{H}(\hat{\mu}) = \kappa\hat{\mu}$, this constraint must bind at the optimum, that is, the optimal choice of κ is

$$\kappa_{3S}^* = \frac{(1 - \mu_0) - (1 - \hat{\mu})^2}{\hat{\mu}(2 - \hat{\mu})}.$$

Proposition 6. *The receiver-optimal three-signal test is*

$$\mathcal{H}_{3S}^*(\mu) = \frac{(1 - \mu_0) - (1 - \hat{\mu})^2}{\hat{\mu}(2 - \hat{\mu})} \mu + \frac{2 - \mu_0 - \hat{\mu}}{2 - \hat{\mu}} (\mu - \hat{\mu})^+,$$

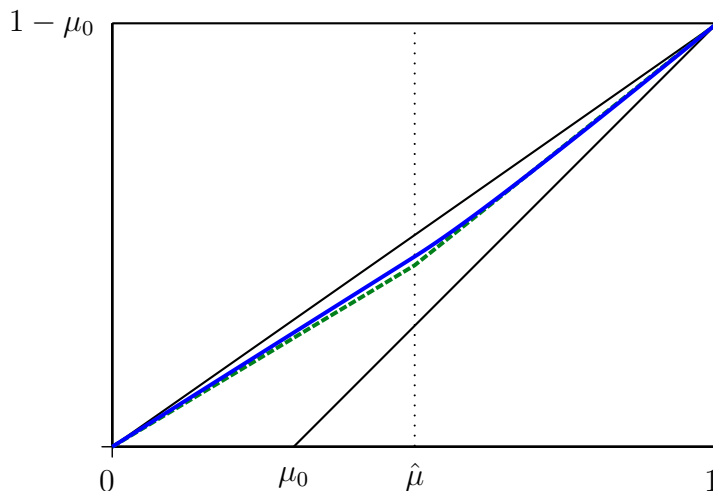*and it corresponds to the one described in Proposition 2.*

24

Figure 8: *Optimal Design – the lower dashed curve is the receiver-optimal three-signal test, and the higher curve is our receiver-optimal test.*

This experiment is illustrated in Figure 8, which also depicts the optimal test that we characterize next. In order to do so, we first define the unique test that makes the persuader indifferent across all falsification levels p_B that induce an approval threshold between $\hat{\mu}$ and 1. Then, we proceed to show that this test is optimal. Such a test must satisfy the incentive constraint (IC$'_0$) everywhere with equality, and must therefore solve the *indifference differential equation*

$$H(\mu) - \frac{\mu - \hat{\mu}}{\mu(\mu - \mu_0)}\mathcal{H}(\mu) = \frac{\kappa\hat{\mu}}{\mu}, \tag{IDE}$$

on $\left[\hat{\mu}, 1\right]$, with initial condition $\mathcal{H}(\hat{\mu}) = \kappa\hat{\mu}$. The unique solution to this problem is given by

$$\mathcal{H}(\mu) = \kappa\hat{\mu}\psi(\mu)\left(1 + \int_{\hat{\mu}}^{\mu} \frac{1}{x\psi(x)}dx\right),$$

where

$$\psi(\mu) = \exp\left(\int_{\hat{\mu}}^{\mu} \frac{x - \hat{\mu}}{x(x - \mu_0)}dx\right).$$

If $\mathcal{H} \in \Delta^B$, it must satisfy $\mathcal{H}(1) = 1 - \mu_0$. Adding this constraint pins down the value of κ to

$$\kappa^* = \frac{1 - \mu_0}{\hat{\mu}\psi(1)\left(1 + \int_{\hat{\mu}}^{1} \frac{1}{x\psi(x)}dx\right)}.$$

25

Theorem 1. *The test defined by*

$$\mathcal{H}^*(\mu) = \begin{cases} \kappa^* \mu & \text{if } \mu \leq \hat{\mu} \\ \kappa^* \hat{\mu} \psi(\mu) \left(1 + \int_{\hat{\mu}}^{\mu} \frac{1}{x\psi(x)} dx \right) & \text{if } \mu \geq \hat{\mu} \end{cases}$$

is optimal. Furthermore, any other optimal test must be linear to the left of $\hat{\mu}$ and less informative than $\mathcal{H}^*$.

Proof. The proof consists of three steps. The first step is to show that $\mathcal{H}^*$ is indeed in Δ^B, so that it is actually a test. This purely calculatory part is proved in the appendix. The third step is to show that any other optimal test is linear to the left of $\hat{\mu}$, and less informative. It is relegated to the appendix as well. In what follows, we provide the second and most interesting step of the proof, which consists in showing that no incentive compatible test can do better than $\mathcal{H}^*$.

To see this, suppose that there exists a test $\mathcal{H} \in \Delta^B$ that satisfies (IC$_0'$), and $\mathcal{H}(\hat{\mu}) > \mathcal{H}^*(\hat{\mu})$. Lemma 4 implies that we can additionally chose it to be linear to the left of $\hat{\mu}$, with slope $\kappa > \kappa^*$, as $\kappa\hat{\mu} = \mathcal{H}(\hat{\mu}) > \mathcal{H}^*(\hat{\mu}) = \kappa^*\hat{\mu}$. Since $\mathcal{H}(1) = \mathcal{H}^*(1) = 1 - \mu_0$, the intermediate value theorem applied to the difference of $\mathcal{H} - \mathcal{H}^*$, which is continuous by convexity of each of these functions, implies that $\mathcal{H}$ and $\mathcal{H}^*$ cross at least once on $(\hat{\mu}, 1]$. Let $\tilde{\mu}$ be the smallest of these crossing points. Then $\mathcal{H}(\mu) > \mathcal{H}^*(\mu)$ for every $\mu \in [\hat{\mu}, \tilde{\mu}]$, which implies that the left-derivative of $\mathcal{H}$ at $\tilde{\mu}$ is smaller than the left derivative of $\mathcal{H}^*$ at $\tilde{\mu}$, that is $H(\tilde{\mu}) \leq H^*(\tilde{\mu})$. Therefore, we have

$$\tilde{\mu} H(\tilde{\mu}) - \frac{\tilde{\mu} - \hat{\mu}}{\tilde{\mu} - \mu_0} \mathcal{H}(\tilde{\mu}) \leq \tilde{\mu} H^*(\tilde{\mu}) - \frac{\tilde{\mu} - \hat{\mu}}{\tilde{\mu} - \mu_0} \mathcal{H}^*(\tilde{\mu}) = \kappa^* \hat{\mu} < \kappa \hat{\mu},$$

which implies that $\mathcal{H}$ cannot satisfy (IC$_0'$), a contradiction. $\qquad\square$

The optimal test is illustrated in Figure 8 and Figure 9. In the proof of Theorem 1, we derive a closed form expression of the optimal test without integrals. For every $\mu \geq \hat{\mu}$,

$$\mathcal{H}^*(\mu) = \kappa^*(\mu - \mu_0) \left\{ 1 + \mu_0 (\hat{\mu} - \mu_0)^{\frac{\hat{\mu}}{\mu_0} - 1} \hat{\mu}^{-\frac{\hat{\mu}}{\mu_0}} \left(\frac{\mu}{\mu - \mu_0} \right)^{\frac{\hat{\mu}}{\mu_0}} \right\}.$$

Using this expression we establish that $\mathcal{H}^*$ satisfies the following properties:

Proposition 7. *The belief distribution generated by the optimal test has support on $\{0\} \cup [\hat{\mu}, 1]$, with atoms at 0 and 1, and a positive, continuously differentiable, and decreasing density on*
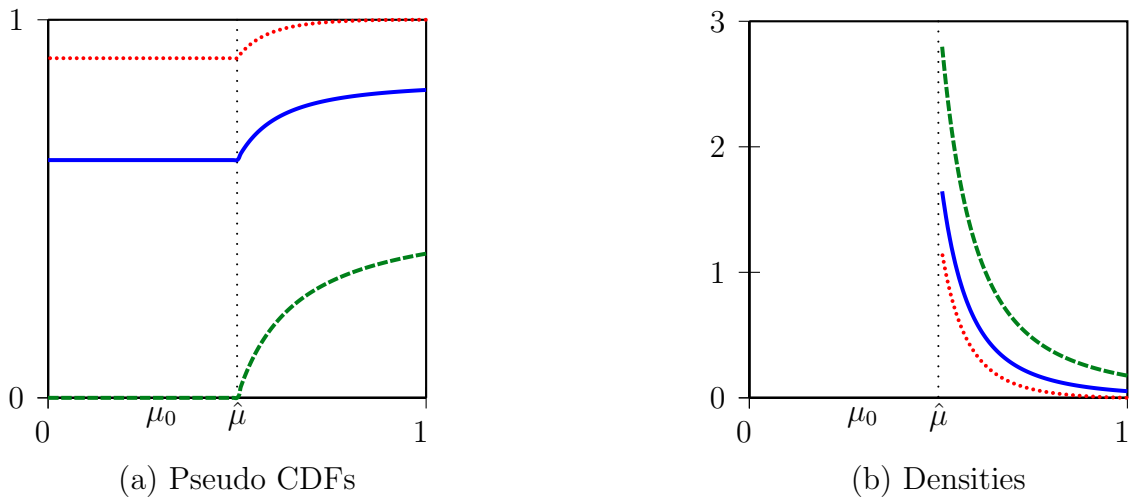
Figure 9: *Optimal Design – in each panel, the blue curve in the middle is the distribution of beliefs, the dashed green curve is the distribution of beliefs generated by the good type, and the dotted red curve is the distribution of beliefs generated by the bad type.*

$[\hat{\mu}, 1)$. The belief distribution generated by the good type has support on $[\hat{\mu}, 1]$, with a positive, continuously differentiable, and decreasing density on $[\hat{\mu}, 1)$, and a single atom at 1. The belief distribution of the bad type has support on $\{0\} \cup [\hat{\mu}, 1]$, with a single atom at 0, and a positive, continuously differentiable, and decreasing density on $[\hat{\mu}, 1)$. Furthermore, the belief distribution generated by the good type first-order stochastically dominates that of the bad type.

Hence, optimal tests use a rich set of signals. They involve 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. Note that Figure 9 shows a clustering of grades close to the threshold. 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 bad type generates the continuum of signals above $\hat{\mu}$ rather than the reject signal. But the reciver would react by rejecting some of the signals above $\hat{\mu}$ in an amount that exactly offsets the advantage from the first effect.

Our optimal test makes the persuader indifferent across all moderate levels of falsification as it satisfies (IDE). Indifference of "the persuader" 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 persuader.

Implementation. As in the three-signal example, there are multiple ways to implement the optimal information structure. Obtaining perfect information and then garbling it before transmitting it to the receiver is one way. Another way is to design a continuum of pass-fail tests assigned to each item randomly and independently with carefully chosen probabilities. Each of these pass-fail tests is failed only by the bad type, but can be passed by both, so that passing leads to a belief $\mu \geq \hat{\mu}$, and these beliefs index the continuum of pass-fail tests. The fully informative pass-fail test is assigned with probability $1 - H(1)$, whereas the other tests are assigned with probability $h_G(\mu)$, and are such that the good type passes with probability 1, but the bad type only with probability $h_B(\mu)/h_G(\mu)$, so that passing leads to belief μ.

Performance. We compare the performance of optimal tests and optimal three-signal tests with full information for the receiver. This comparison is meant as a simple illustration and it is depicted in Figure 10 which also gives a sense of comparative statics. Both optimal tests deliver at least 50% of the full information payoff. A numerical analysis shows that the optimal three-signal test delivers at least around 80% of the optimal test suggesting that most of the benefits can be harvested with simple tests using a small number of signals.

Proposition 8. $\mathcal{H}^*$ and $\mathcal{H}^*_{3S}$ are ex-ante Pareto efficient. With both tests, the receiver obtains at least 1/2 of the full information payoff. Furthermore, this bound is strict since one can find a sequence of pairs $(\mu_0, \hat{\mu})$ such that the payoff ratio gets arbitrarily close to 1/2.

Figure 11 shows the outcome of different information structures in the payoff space, and illustrates the efficiency of both tests. The outcome is always on the Pareto frontier.

8   Costly Falsification

In this section, we study receiver-optimal test design when falsification is costly. We model this with a cost function $C(p_B, p_G) \geq 0$. The cost can be thought of as a combination of a technological scaling cost, and an expected punishment cost of being caught-which could be explicit, psychological, or reputational. We naturally assume that $C(\cdot)$ is continuous and increasing in p_B and p_G, and that $C(0,0) = 0$. The optimal approval strategy described in

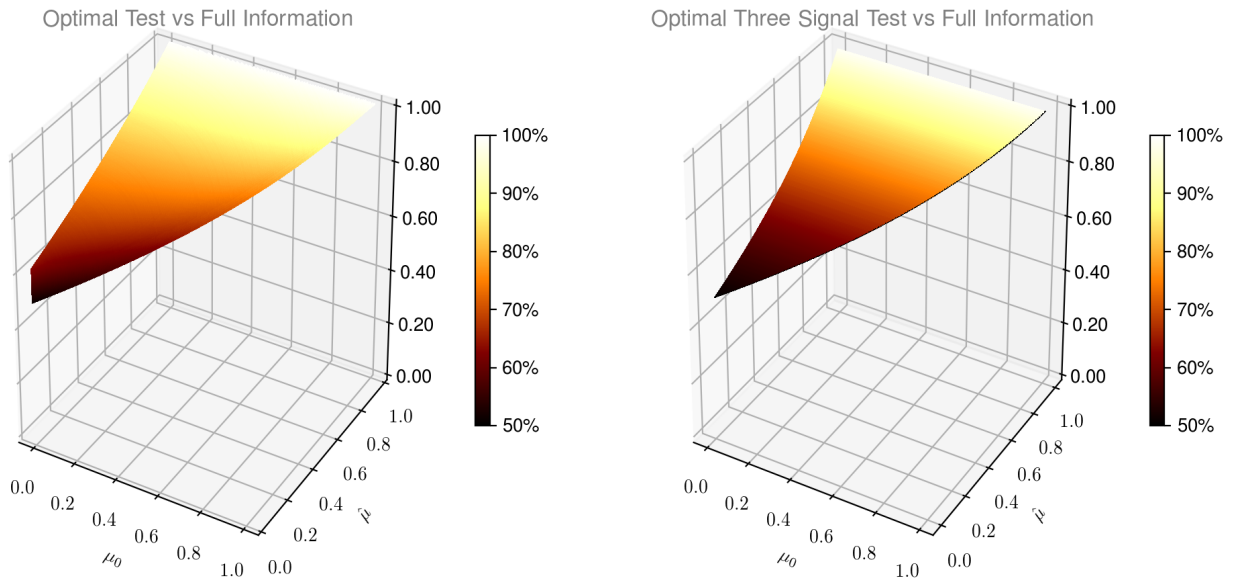**Figure 10:** *Performance of $\mathcal{H}^*$ and $\mathcal{H}_{3S}^*$ in percentage of the full information payoff*
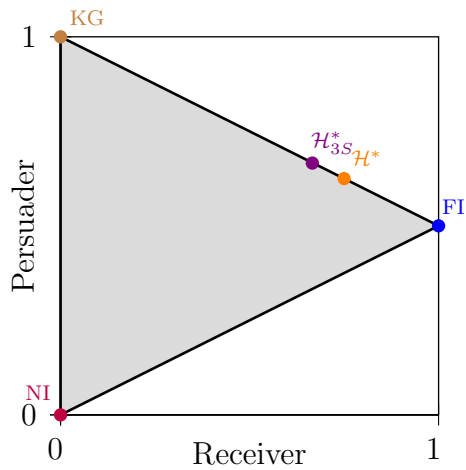

Figure 11: *Information structures in payoff space. Each player's payoff is expressed in percentage of her maximum attainable payoff. The grey triangle is the space of attainable payoffs, and the dots represent the payoffs achieved by different information structures.*

Proposition 3 applies to the case of costly falsification without any modifications. Then, the fact that $C(p_B, p_G)$ is increasing in p_G ensures that the optimal falsification result of Proposition 4 holds with cost, so the persuader always chooses $p_G = 0$. Furthermore, the relevant range for p_B is again the interval $I = \left[0, \frac{\mu_0(1-\hat{\mu})}{\hat{\mu}(1-\mu_0)}\right]$. As a consequence, to simplify notation, we can define the new cost function $c(p_B) = C(p_B, 0)$.

An important building block of our analysis is the no-falsification principle. In order for the principle to hold, it must be no more costly to raise falsification from any p_B^* to $p_B^* + (1-p_B^*)\varepsilon$, than it is to raise it from 0 to ε. This is satisfied whenever $c(p_B)$ is concave in p_B, but we can also accommodate some moderately convex functions with a positive marginal cost at 0. The following assumption on the cost function ensures that the no-falsification principle holds.[12]

Assumption 3. *For every $p_B \in I$ and every $\varepsilon > 0$ such that $p_B + \varepsilon \in I$,*

$$c(\varepsilon) \geq c\big(p_B + (1-p_B)\varepsilon\big) - c(p_B).$$

Under Assumption 3, we can formulate the optimal design problem as before. The only difference is that we need to account for the cost in the no-falsification incentive constraint, which becomes

$$\frac{\mu - \hat{\mu}}{\mu - \mu_0}\mathcal{H}(\mu) - \hat{\mu}c\left(\frac{\mu_0(\mu - \hat{\mu})}{\hat{\mu}(\mu - \mu_0)}\right) \leq \mu H(\mu) - \hat{\mu}H(\hat{\mu}), \qquad \forall \mu \in \big[\hat{\mu}, 1\big]. \tag{IC$_0^c$}$$

Intuitively, costly falsification should allow us to attain more informative information structures. Hence, we can start by looking for conditions on the cost function that allow us to attain full information. The fully informative test is given by $\mathcal{H}(\mu) = (1 - \mu_0)\mu$, and is incentive compatible if, for every $\mu \in \big[\hat{\mu}, 1\big]$,

$$c\left(\frac{\mu_0(\mu - \hat{\mu})}{\hat{\mu}(\mu - \mu_0)}\right) \geq (1 - \mu_0)\frac{\mu_0(\mu - \hat{\mu})}{\hat{\mu}(\mu - \mu_0)}.$$

That is, if the cost function satisfies the following full information condition

$$c(p_B) \geq (1 - \mu_0)p_B, \qquad \forall p_B \in I. \tag{FI}$$

---

[12]Note that, if $c(\cdot)$ is differentiable at 0, Assumption 3 is equivalent to requesting that $c'(0) \geq (1 - p_B)c'(p_B)$ for every $p_B \in I$ at which $c(\cdot)$ is differentiable.

This also shows (replacing the inequality by an equality), that the cost function $c(p_B) = (1 - \mu_0)p_B$ is the unique one that makes the persuader indifferent across all the thresholds he might induce by falsifying under the fully informative test.

In what follows, we assume that $c(p_B) = \lambda p_B$, with $\lambda > 0$. Such linear cost functions lend themselves to interesting comparative static results and tractable analysis.[13] Note that Assumption 3 is automatically satisfied by linear cost functions. Moreover, $c(p_B)$ satisfies (FI) if and only if $\lambda \geq 1 - \mu_0$. Otherwise, we write the indifference differential equation, which is given by

$$H(\mu) - \frac{\mu - \hat{\mu}}{\mu(\mu - \mu_0)}\mathcal{H}(\mu) = \frac{\kappa\hat{\mu}}{\mu} - \lambda\frac{\mu_0(\mu - \hat{\mu})}{\mu(\mu - \mu_0)}.$$

Its solution with initial condition $\mathcal{H}(\hat{\mu}) = \kappa\hat{\mu}$ is

$$\mathcal{H}(\mu) = \hat{\mu}\psi(\mu)\left[\kappa\left(1 + \int_{\hat{\mu}}^{\mu}\frac{1}{x\psi(x)}dx\right) - \lambda\frac{\mu_0}{\hat{\mu}}\int_{\hat{\mu}}^{\mu}\frac{x - \hat{\mu}}{x(x - \mu_0)\psi(x)}dx\right],$$

and the unique value of κ that ensures that $\mathcal{H}(1) = 1 - \mu_0$ is

$$\kappa_\lambda^* = \left(\frac{1 - \mu_0}{\hat{\mu}\psi(1)} + \lambda\frac{\mu_0}{\hat{\mu}}\int_{\hat{\mu}}^{1}\frac{x - \hat{\mu}}{x(x - \mu_0)\psi(x)}dx\right)\left(1 + \int_{\hat{\mu}}^{1}\frac{1}{x\psi(x)}dx\right)^{-1}.$$

Then, we have the following result.

Theorem 2. *If $\lambda \geq 1 - \mu_0$, then the optimal test is the fully informative one. Otherwise, the test given by*

$$\mathcal{H}_\lambda^*(\mu) = \begin{cases} \kappa_\lambda^*\mu & \text{if } \mu \leq \hat{\mu} \\ \hat{\mu}\psi(\mu)\left[\kappa_\lambda^*\left(1 + \int_{\hat{\mu}}^{\mu}\frac{1}{x\psi(x)}dx\right) - \lambda\frac{\mu_0}{\hat{\mu}}\int_{\hat{\mu}}^{\mu}\frac{x - \hat{\mu}}{x(x - \mu_0)\psi(x)}dx\right] & \text{if } \mu \geq \hat{\mu} \end{cases}$$

is optimal. Furthermore, any other optimal test must be linear to the left of $\hat{\mu}$, and less informative than $\mathcal{H}_\lambda^*$. Finally, for all $\mu \in (0,1)$, $\mathcal{H}_{FI}(\mu) > \mathcal{H}_\lambda^*(\mu) > \mathcal{H}^*(\mu)$.

---

[13] The complete solution for arbitrary cost functions that satisfy Assumption 3 is complicated because the solution of the differential equation may not define a test. In Appendix C, we show how we can modify the cost function recursively to obtain a solution for a more general class of cost functions. In the case of a linear cost, the recursive approach is not necessary.

In the proof of [Theorem 2](), we derive the following expression for $\mathcal{H}_\lambda^*$. For every $\mu \geq \hat{\mu}$,

$$\mathcal{H}_\lambda^*(\mu) = \kappa_\lambda^* \mu + (\kappa_\lambda^* - \lambda)\mu_0 \left\{ \left(\frac{\mu}{\hat{\mu}}\right)^{\frac{\hat{\mu}}{\mu_0}} \left(\frac{\hat{\mu} - \mu_0}{\mu - \mu_0}\right)^{\frac{\hat{\mu}}{\mu_0} - 1} - 1 \right\}.$$

With a linear cost, the optimal test has the same qualitative properties as without cost.

Proposition 9. *Suppose $\lambda < 1 - \mu_0$. Then, the belief distribution generated by our optimal test has support on $\{0\} \cup [\hat{\mu}, 1]$, with atoms at 0 and 1, and a positive, continuously differentiable, and decreasing density on $[\hat{\mu}, 1)$. The belief distribution generated by the good type has support on $[\hat{\mu}, 1]$, with a positive, continuously differentiable, and decreasing density on $[\hat{\mu}, 1)$, and a single atom at 1. The belief distribution of the bad type has support on $\{0\} \cup [\hat{\mu}, 1]$, with a single atom at 0, and a positive, continuously differentiable, and decreasing density on $[\hat{\mu}, 1)$. Furthermore, the belief distribution generated by the good type first-order stochastically dominates that of the bad type.*

In addition, we can derive the following comparative statics in λ confirming the initial intuition that higher costs lead to more informative optimal tests.

Proposition 10. *For $\lambda \leq 1 - \mu_0$, the Blackwell informativeness of $\mathcal{H}_\lambda^*$ is strictly increasing in λ.*

9 Relaxing perfect observability and falsification limits.

In the baseline analysis, we have assumed that falsification rates are perfectly observable by the receiver ([Assumption 1]()), and that they must satisfy $p_B + p_G \leq 1$ ([Assumption 2]()). The latter assumption guarantees that the meaning of grades is not flipped (higher signals are associated with a higher belief that an item is good). Interestingly, as we explain in [Appendix B](), the reason we need [Assumption 2]() is because we impose the perfect observability [Assumption 1](). However, perfect observability is likely to be unjustified in many contexts. We now drop both these assumptions and derive the optimal falsification-proof test in the limit case where the persuader has a continuum of IID items up for approval. We also sketch how these assumptions can be relaxed in a model of sequential decisions.

9.1 Continuum of Items

On the equilibrium path falsification rates are correctly anticipated even if they are unobserved. The issue arises for off-path information sets. Below we tackle the issue of off-path information sets and explain why the test we derive in the main analysis remains optimal. The main intuition is as follows. When perfect observability is relaxed, the receiver can still partially infer manipulation behavior from the cross-sectional distribution of signals. We show that, as long as falsification is costly, among all falsification rates that generate the same information set for the receiver, one strictly dominates all the other. Therefore, in a subgame perfect equilibrium, conditional on reaching a certain information set, the receiver knows for sure what choice the persuader must have made, and can adopt the same beliefs as in the case of perfect observability. This is true for information sets both on and off the equilibrium path. Therefore, all results in the costly case still hold when the auxiliary assumptions are relaxed.

For the costless case, they extend through two arguments. The first one is a selection argument. By taking a falsification cost that converges to 0, we obtain our optimal test in the costless case. The second argument relies on the idea that the persuader, conditional on attaining any given payoff, should prefer lower falsification rates. This can be nicely captured by assuming that the persuader has lexicographic preferences, with approval rate as its first dimension, and any decreasing function of p_B, and p_G on the second dimension. Under such lexicographic preferences, the dominance argument holds as well, implying that our optimal test in the costless case is optimal in this relaxed setup as well.

Exploiting the Empirical Distribution of Test Results. Since the persuader has a continuum of IID items that he subjects to testing, the receiver can make inferences about the persuader's falsification rates from the empirical distribution of test results:[14]

Given a test H, for any choice of falsification (p_B, p_G), the cross-sectional distribution of

---

[14]Such linking of decisions has shown to be useful by Jackson and Sonnenschein (2007) who establish that the incentive costs become negligible by constructing a mechanism in which each persuader announces preferences over many decisions. These announcements must be "budgeted" such that the distribution of types across problems must mirror the underlying distribution of their preferences. Analogously, in our setup Bayes' rule implies the distribution of posteriors must integrate to the prior.
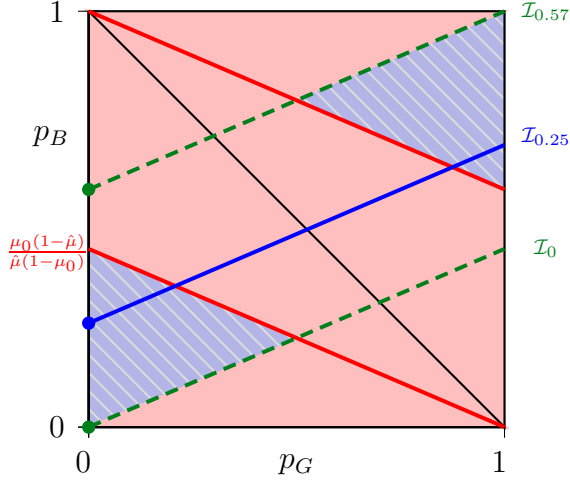
Figure 12: *The blue line, and the green dashed lines each depict an information set of the receiver, that is a set of falsification rates that she cannot tell apart. On each of these information sets, the dot shows the only undominated strategy (p_B^α, p_G^α) of the persuader.*

signals observed by the receiver is

$$
F(\mu) = \left\{\mu_0(1 - p_G) + (1 - \mu_0)p_B\right\}H_G(\mu) + \left\{\mu_0 p_G + (1 - \mu_0)(1 - p_B)\right\}H_B(\mu)
$$
$$
= H(\mu) + \left(\frac{p_G}{1 - \mu_0} - \frac{p_B}{\mu_0}\right)\left\{\mathcal{H}(\mu) - (\mu - \mu_0)H(\mu)\right\}.
$$

Hence, for every test that is not the uninformative test, the receiver can compute $\frac{p_G}{1-\mu_0} - \frac{p_B}{\mu_0}$ from the cross-sectional distribution of signals. She cannot perfectly observe the choice of falsification of the persuader, since she cannot tell apart two strategies (p_B, p_G) and (p'_B, p'_G) such that $\frac{p_G}{1-\mu_0} - \frac{p_B}{\mu_0} = \frac{p'_G}{1-\mu_0} - \frac{p'_B}{\mu_0}$. Therefore, the information sets of the receiver are the sets

$$
\mathcal{I}_\alpha = \left\{(p_B, p_G) \in [0, 1]^2 : p_B = \frac{\mu_0}{1 - \mu_0}p_G + \alpha\right\},
$$

for $\alpha \in [-1, 1]$.

A strategy of the receiver specifies an approval policy conditioned on signals for each of her information sets. Since all falsification choices (p_B, p_G) that belong to the same information set $\mathcal{I}_\alpha$ generate the same distribution of signals F, any strategy of the receiver leads to the same approval probabilities of good and bad items for all $(p_B, p_G) \in \mathcal{I}_\alpha$. When falsification is

costless, the persuader is thus indifferent between any two falsification strategies in the same information set. However, when there are even mild falsification costs which increase with the levels of falsification, this indifference breaks down. We discuss this case first.

Whenever falsification is costly, as in Section 8, with a cost function $C(p_B, p_G) \geq 0$ that is increasing, any strategy $(p_B, p_G) \in \mathcal{I}_\alpha$ that does not minimize p_G (and p_B) is strictly dominated by the one that minimizes falsification rates, and thus associated costs,

$$(p_B^\alpha, p_G^\alpha) = \min \ \mathcal{I}_\alpha.$$

The cost-minimizing falsification strategies $\left\{(p_B^\alpha, p_G^\alpha)\right\}_{\alpha \in [-1,1]}$ all satisfy $p_B^\alpha + p_G^\alpha \leq 1$. Furthermore, they contain all falsification strategies of the form $(p_B, 0)$ with $p_B \leq \frac{\mu_0(1-\hat{\mu})}{\hat{\mu}(1-\mu_0)}$, that is all the falsification choices that were potentially optimal in our former analysis (see Proposition 4).

Falsification strategies that do not belong to $\left\{(p_B^\alpha, p_G^\alpha)\right\}_{\alpha \in [-1,1]}$ are strictly dominated and cannot be equilibrium strategies. Therefore, when reaching information set $\mathcal{I}_\alpha$, the receiver's equilibrium belief must be, accurately, that the persuader played (p_B^α, p_G^α). Hence, our analysis of costly falsification (Section 8 and Appendix C) carries on to the case where Assumption 1 and Assumption 2 are relaxed, and all results hold. In particular, the problem of finding an optimal test can be reduced to maximizing $\mathcal{H}(\hat{\mu})$ over test functions $\mathcal{H} \in \Delta^B$ under the constraint (IC_0^c).

To extend our results in the costless case, we can follow two routes. The first option is a selection argument which consists of looking at the limit of the costly falsification problem with a vanishing cost. Consider the (linear) cost function $\varepsilon C_\lambda(p_B, p_G)$, where $C_\lambda(p_B, 0) = \lambda p_B$. Then, the following result is immediate:

Proposition 11. *The test $\mathcal{H}_{\varepsilon\lambda}^*$ is optimal under the cost function $\varepsilon C_\lambda(p_B, p_G)$, and it uniformly converges to $\mathcal{H}^*$ as $\varepsilon \to 0$.*

The second option, is to consider a persuader with lexicographic preferences with approval probability as the first dimension, and an increasing falsification cost as the second dimension. Such preferences naturally capture a distaste for falsification at a given payoff level. The strict domination argument we made is still valid with these lexicographic preferences, and therefore the rest of the analysis follows as well, leading to the following result:

Theorem 3. *Under lexicographic preferences with any increasing cost function, the test $\mathcal{H}^*$ is receiver-optimal.*

9.2 Dynamic Interaction

Our optimal test in the static model remains optimal in a dynamic, and in some cases more realistic, scenario that does not rely on Assumption 1 and Assumption 2. Suppose that time is discrete and there is an infinite number of periods. In period zero the persuader, faced with a test, chooses his falsification rates (p_G, p_B) which remain unchanged throughout. Falsification rates are unobserved, but correctly anticipated in equilibrium. In each subsequent period, an item is tested, and the result becomes public. The receiver (or the period-t receiver[15]), then decides whether or not to approve the item based on its test result having access to all past histories of test results. Suppose that the test the optimal one we derived and for simplicity suppose that the persuader can only choose p_B.

We now sketch why in equilibrium the persuader optimally chooses $p_B = 0$. To establish whether or not $p_B = 0$ is a best response, we have to evaluate what is the persuader's payoff if he deviates. Now such a deviation is unobservable. However, the falsification rate will eventually become apparent from the empirical history of test results. To be able to handle technicalities that arise from the need to update about the likelihood of falsification rates that occur with probability zero on the equilibrium path, we can rely on trembling-hand equilibrium. The trembles here imply that each falsification rate $p_B \neq 0$ arises with some positive but, possibly, arbitrarily small probability. Formally the persuader chooses the intended falsification rate p_B with probability $1 - \varepsilon$ and all remaining falsification rates with probability ε. In other words, it is as if he chooses the dirac measure δ_{p_B} with probability $1 - \varepsilon$ and the uniform distribution on $[0, 1]$ with probability ε. This gives a consistent way with which a receiver at period t evaluates a history of realized results: $\mu^t = (\mu_1, \ldots, \mu_{t-1})$. As the time progresses receivers start assigning more and more weight on the actual falsification rate chosen. When the history is long enough, standard results in Bayesian statistics imply that the receiver will put almost all weight on the actual choice of p_B.

For each history there is an updated distribution of the falsification rate, $\lambda(p_B|\mu^t)$ and the receiver approves based on that. Given that $\lambda(p_B|\mu^t)$ converges to the truth at $t \to \infty$, (which is a degenerate distribution with a point mass on the falsification rate actually chosen by the persuader) the receivers will behave as in the model where the p_B is known (as when

---

[15] The receiver can also be taken to be a sequence of identical short-lived receivers. The principal and the persuader assign equal weights on each period's payoff.

Assumption 1 holds). Then, given the proposed test we know that choosing $p_B = 0$ is a best response, and in fact the same argument shows that the persuader's best response to any test in the static setup remains a best response in the dynamic one. Hence the optimal test we derived remains optimal in the dynamic setup.

10   Concluding Remarks

We study optimal tests in the presence of falsification. Our results deliver insights for how to enhance the reliability of tests that persuaders 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 persuader: Any binary test can be turned to deliver the persuader-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 persuader. 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 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 persuader knows the types of his item(s) (interim falsification). Second, we can accommodate multiple persuaders, 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 falsi-

fier 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 persuader knows the true μ_0. Then using our optimal test for a particular value μ_0' would lead each persuader with a different realization μ_0 to falsify so as to generate the same grade distribution as a persuader with μ_0' and no falsification. So a persuader with $\mu_0 > \mu_0'$ would set $p_G > 0$, and a persuader 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 persuader exerts. If production costs are sufficiently low, then the persuader 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-persuader 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

A   Proofs without Cost

Proof of Lemma 1. Let $\mathcal{H} \in \Delta^B$. By convexity, $\mathcal{H}$ has a left derivative everywhere on $(0, 1]$, let $H(\mu)$ be the left derivative of $\mathcal{H}$ at μ. Furthermore, H is piecewise-continuous, everywhere left-continuous, and weakly increasing on $(0, 1]$. Then, we can define $H(0) = \lim_{\mu \to 0} H(\mu)$. Because, $\mathcal{H}$ is increasing, H is non-negative. It is also bounded above by 1. Suppose not, so that $H(\mu) > 1$ for some μ. Because H is left-continuous, there must be an interval $[\mu - \varepsilon, \mu]$ to

the left of μ such that $H(x) > 1$ for all $x \in [\mu - \varepsilon, \mu]$, so we can choose $x < 1$ such that $H(x) > 1$. By convexity, we must have $\mathcal{H}(1) - \mathcal{H}(x) \geq H(x)(1 - x) > 1 - x$. Since $\mathcal{H}(1) = 1 - \mu_0$, this implies $\mathcal{H}(x) < x - \mu_0$, but then $\mathcal{H}$ would violate the lower bound condition on Δ^B.

Next, let H be a probability measure on $[0, 1]$ with mean μ_0, and also the associated pseudo cdf. Define $\mathcal{H}(\mu) = \int_0^\mu H(x)dx$. This function is increasing since H is nonnegative. It is also convex as the integral of a non-decreasing function. The condition on the mean implies that $\mathcal{H}(1) = 1 - \mu_0$, and $\mathcal{H}(0) = 0$ by definition. Suppose that for some $x \in (0, 1)$, $\mathcal{H}(x) > x(1 - \mu_0)$. Then, convexity of $\mathcal{H}$ would imply that

$$\mathcal{H}(1) \geq \mathcal{H}(x) + \frac{\mathcal{H}(x) - \mathcal{H}(0)}{x}(1 - x) = \frac{\mathcal{H}(x)}{x} > 1 - \mu_0,$$

a contradiction. Similarly, if for some $x > \mu_0$, we had $\mathcal{H}(x) < (x - \mu_0)^+$, convexity would imply that $\mathcal{H}(0) < 0$, a contradiction. $\qquad\square$

Proof of Lemma 3. Let $\lambda(\mu) \equiv \frac{H_G(d\mu)}{H_B(d\mu)}$ denote the likelihood ratio induced by the test when the signal realization (and the belief in the absence of falsification) is a small interval $d\mu$ centered on μ. In the presence of falsification, the signal μ observed as a result of the test can no longer be identified with the the belief formed by the receiver. Specifically, by Bayes rule, the belief $\tilde{\mu}$ that is formed when signal μ is generated satisfies

$$\tilde{\mu} = \frac{\mu_0 \tilde{\lambda}(\mu)}{\mu_0 \tilde{\lambda}(\mu) + 1 - \mu_0}, \tag{3}$$

where
$$\tilde{\lambda}(\mu) = \frac{F_G(d\mu)}{F_B(d\mu)} = \frac{(1 - p_G)H_G(d\mu) + p_G H_B(d\mu)}{(1 - p_B)H_B(d\mu) + p_B H_G(d\mu)} = \frac{(1 - p_G)\lambda(\mu) + p_G}{p_B \lambda(\mu) + 1 - p_B}$$

is the new relevant likelihood ratio. This expression is increasing in λ over $[0, \infty)$ whenever $p_B + p_G < 1$, meaning that the post-falsification belief is increasing in the initial belief. By contrast, if $p_B + p_G > 1$, it is decreasing in λ. This relationship can be inverted to get

$$\lambda(\mu) = \frac{(1 - p_B)\tilde{\lambda}(\mu) - p_G}{1 - p_G - p_B \tilde{\lambda}(\mu)}.$$

A simple rewriting of (3) also gives us: $\tilde{\lambda}(\mu) = \frac{\tilde{\mu}(1 - \mu_0)}{\mu_0(1 - \tilde{\mu})}$. Using these expressions, we can write

the signal, and original belief, as a function of the post-falsification belief:

$$\mu = \frac{\mu_0}{\mu_0 + (1 - \mu_0)\lambda(\mu)^{-1}}$$

$$= \frac{\mu_0}{\mu_0 + (1 - \mu_0)\frac{1 - p_G - p_B\tilde{\lambda}(\mu)}{(1 - p_B)\tilde{\lambda}(\mu) - p_G}}$$

$$= \frac{\mu_0}{\mu_0 + (1 - \mu_0)\frac{1 - p_G - p_B\frac{1 - \mu_0}{\mu_0}\frac{\tilde{\mu}}{1 - \tilde{\mu}}}{(1 - p_B)\frac{1 - \mu_0}{\mu_0}\frac{\tilde{\mu}}{1 - \tilde{\mu}} - p_G}}$$

$$= \frac{\mu_0}{\mu_0 + (1 - \mu_0)\frac{\mu_0(1 - p_G) - \tilde{\mu}(p_B + \mu_0(1 - p_G - p_B))}{\tilde{\mu}(1 - p_B - \mu_0(1 - p_B - p_G)) - \mu_0 p_G}}$$

$$= \frac{\mu_0\left(\tilde{\mu}\left(1 - p_B - \mu_0(1 - p_B - p_G)\right) - \mu_0 p_G\right)}{\mu_0(1 - p_G) - \tilde{\mu}\left(p_B + \mu_0(1 - p_G - p_B)\right) + \mu_0(\tilde{\mu} - \mu_0)}$$

$$= \frac{\mu_0\tilde{\mu}\left(1 - p_B - \mu_0(1 - p_B - p_G)\right) - \mu_0^2 p_G}{\tilde{\mu}\left(\mu_0(p_B + p_G) - p_B\right) + \mu_0(1 - \mu_0) - \mu_0 p_G}$$

$$= \mu_0\frac{(1 - \mu_0)\tilde{\mu} - \mu_0(1 - \tilde{\mu})p_G - (1 - \mu_0)\tilde{\mu}p_B}{\mu_0(1 - \mu_0) - \mu_0(1 - \tilde{\mu})p_G - (1 - \mu_0)\tilde{\mu}p_B}.$$

It is easy to see that $\tilde{\mu}$ lies in $\left[\frac{\mu_0 p_G}{\mu_0 p_G + (1 - \mu_0)(1 - p_B)}, \frac{\mu_0(1 - p_G)}{\mu_0(1 - p_G) + (1 - \mu_0)p_B}\right]$. The remaining points follow from easy calculations. $\qquad\square$

Proof of Lemma 2. We show the proof for F_G, it is similar for F_B. Consider the joint probability that a certain item is of the good type, and the information structure generates a belief in $[0, \mu)$ for this item. This probability can be written as $\mu_0 F_G(\mu)$, or as $\int_0^\mu x F(dx)$. By integration by parts, the latter is equal to $\mu F(\mu) - \mathcal{F}(\mu)$, which concludes the proof. $\qquad\square$

Proof of Proposition 3. If $p_B + p_G = 1$, the resulting information structure is uninformative, the receiver has belief μ_0 regardless of the signal and does not approve. Next, we treat the case $p_B + p_G < 1$. Because μ_0 is the prior, it must lie in the interval $\left[\underline{\mu}, \overline{\mu}\right]$. $\hat{\mu}$, however, need not lie in this interval, and, if it does not, the receiver never approves. This is the case if the upper bound of the interval is below $\hat{\mu}$, that is

$$\frac{\mu_0(1 - p_G)}{\mu_0(1 - p_G) + (1 - \mu_0)p_B} < \hat{\mu} \quad \Leftrightarrow \quad p_B > \frac{\mu_0(1 - \hat{\mu})}{(1 - \mu_0)\hat{\mu}}(1 - p_G).$$

When this is not the case, the receiver approves for beliefs above $\hat{\mu}$, that is for signals above

$$\hat{\mu}(p_B, p_G) = \frac{\mu_0\hat{\mu}\big(1 - p_B - \mu_0(1 - p_B - p_G)\big) - \mu_0^2 p_G}{\hat{\mu}\big(\mu_0(p_B + p_G) - p_B\big) + \mu_0(1 - \mu_0) - \mu_0 p_G}$$
$$= \mu_0 \frac{(1 - \mu_0)\hat{\mu} - \mu_0(1 - \hat{\mu})p_G - (1 - \mu_0)\hat{\mu}p_B}{\mu_0(1 - \mu_0) - \mu_0(1 - \hat{\mu})p_G - (1 - \mu_0)\hat{\mu}p_B}.$$

A simple calculation shows that this $\hat{\mu}(p_B, p_G)$ increases with p_B and p_G for $p_B + p_G < 1$.

Finally, consider the case $p_B + p_G > 1$. Then, the belief transformation is decreasing, and the receiver will therefore approve when signals are below $\hat{\mu}(p_B, p_G)$. As previously, $\hat{\mu}$ may not lie in the interval $\big[\underline{\mu}, \overline{\mu}\big]$. Now, it is the case if $\hat{\mu}$ lies below $\underline{\mu}$, that is

$$\frac{\mu_0 p_G}{\mu_0 p_G + (1 - \mu_0)(1 - p_B)} > \hat{\mu} \quad \Leftrightarrow \quad p_B > 1 - \frac{\mu_0(1 - \hat{\mu})}{(1 - \mu_0)\hat{\mu}}p_G.$$

A simple calculation shows that $\hat{\mu}(p_B, p_G)$ decreases with p_B and p_G for $p_B + p_G > 1$. □

To prove Proposition 4 we need the help of the following lemma.

Lemma 5. *For every $\mu \in [\mu_0, 1]$, $\mathcal{H}(\mu) - (\mu - \mu_0)H(\mu) \geq 0$, and the inequality is strict if and only if $H(\mu) < 1$. Furthermore, this expression is nonincreasing in μ.*

Proof. Since $H(\mu) \leq 1$, we have $\mathcal{H}(\mu) - (\mu - \mu_0)H(\mu) \geq \mathcal{H}(\mu) - (\mu - \mu_0) \geq 0$ by definition of Δ^B, since $(\mu - \mu_0)^+$ is the lower bound of Δ^B. The first inequality is strict if $H(\mu) < 1$. Then, note that, for any $\mu > \mu' > \hat{\mu}$, we have, by convexity

$$\mathcal{H}(\mu) - (\mu - \mu_0)H(\mu) - \mathcal{H}(\mu') + (\mu' - \mu_0)H(\mu') \leq H(\mu')(\mu - \mu') - (\mu - \mu_0)H(\mu) + (\mu' - \mu_0)H(\mu')$$
$$\leq \big(H(\mu') - H(\mu)\big)(\mu - \mu_0) \leq 0$$

□

Proof of Proposition 4. If $H(\hat{\mu}) = 1$, then $H\big(\hat{\mu}(p_B, p_G)\big) = 1$ for any falsification strategy. Therefore, the first term in the expression of $\Pi(p_B, p_G)$ is null, and, by Lemma 5, so is the second term. Hence the payoff of the persuader is null, regardless of her falsification strategy. Furthermore, the receiver approves with probability 0, and therefore her payoff is null.

If $H(\hat{\mu}) < 1$, then no falsification gives the persuader a strictly positive payoff. Therefore any optimal falsification must be such that $p_B \leq \frac{\mu_0(1-\hat{\mu})}{\hat{\mu}(1-\mu_0)}(1 - p_G)$, that is, it must lie below the

41

red line in Figure 5. In addition, it must satisfy $H\big(\hat{\mu}(p_B, p_G)\big) < 1$ and $p_B \geq \frac{\mu_0}{1-\mu_0}p_G$. The second inequality corresponds to the region above the dashed green line in Figure 5. Indeed, a falsification strategy such that $H\big(\hat{\mu}(p_B, p_G)\big) = 1$ would yield a null payoff, and we know that the persuader can do better. Then at any potentially optimal falsification strategy, we have $\mathcal{H}\big(\hat{\mu}(p_B, p_G)\big) - \big(\hat{\mu}(p_B, p_G) - \mu_0\big)H\big(\hat{\mu}(p_B, p_G)\big) > 0$ by Lemma 5. Suppose that $p_B < \frac{\mu_0}{1-\mu_0}p_G$. Then we would have

$$\Pi(p_B, p_G) < 1 - H\big(\hat{\mu}(p_B, p_G)\big) \leq 1 - H(\hat{\mu}),$$

so the persuader would be better off by not falsifying.

Next, let (p_B, p_G) be a falsification strategy that satisfies all these criteria, so that it is potentially optimal. Then $\Pi(p_B, p_G)$ is decreasing in p_G. Indeed, the first term, $1 - H\big(\hat{\mu}(p_B, p_G)\big)$, is nonincreasing in p_G since $\hat{\mu}(p_B, p_G)$ is nondecreasing in p_G. Then $\mathcal{H}\big(\hat{\mu}(p_B, p_G)\big) - \big(\hat{\mu}(p_B, p_G) - \mu_0\big)H\big(\hat{\mu}(p_B, p_G)\big) > 0$ is nonincreasing in p_G by Lemma 5, and $\frac{p_B}{\mu_0} - \frac{p_G}{1-\mu_0} > 0$ is decreasing in p_G. $\qquad\square$

Proof of Proposition 6. We have already proved optimality, so the only thing that remains to be proved is that this experiment indeed corresponds to the one we identified in Proposition 2, that is they generate the same belief distributions. The first experiment generates probability $(1 - \mu_0)(1 - \pi_B^*)$ on 0, and the following calculation shows that this is equal to $H(0) = \kappa$,

$$(1 - \mu_0)(1 - \pi_B^*) = 1 - \mu_0 - \frac{\mu_0(1 - \hat{\mu})^2}{\hat{\mu}(2 - \hat{\mu})}$$
$$= \frac{(1 - \mu_0) - (1 - \hat{\mu})^2}{\hat{\mu}(2 - \hat{\mu})},$$

which concludes the proof since other probabilities must coincide as well for both experiments to generate an average belief of μ_0 and have the same atoms. $\qquad\square$

Proof of Theorem 1. Here, we prove the missing steps in the proof of the theorem.

Step 1. The first step is to prove that $\mathcal{H}^*$ is indeed in Δ^B. Note that $\mathcal{H}^*$ is continuously differentiable, and to show that it is in Δ^B, it is sufficient to show that its derivative H^* is indeed a pseudo cdf. Hence, we show that H^* is nondecreasing and bounded between 0 and 1.

First, note that κ^* is positive. Therefore $H^*(\mu)$ is positive for $\mu \leq \hat{\mu}$. For $\mu > \hat{\mu}$, we know that

$$H^*(\mu) = \frac{\kappa^* \hat{\mu}}{\mu} + \frac{\mu - \hat{\mu}}{\mu(\mu - \mu_0)} \mathcal{H}^*(\mu),$$

and, since $\mathcal{H}^*(\mu)$ is clearly positive, so is $H^*(\mu)$.

Next, we show that H^* is non-decreasing. This is immediate on $[0, \hat{\mu}]$. For $\mu \geq \hat{\mu}$, we start by calculating the integral in the expression of $\psi(\mu)$

$$
\begin{aligned}
\log\big(\psi(\mu)\big) = \int_{\hat{\mu}}^{\mu} \frac{x - \hat{\mu}}{x(x - \mu_0)} dx &= \int_{\hat{\mu}}^{\mu} \frac{1}{x - \mu_0} dx - \int_{\hat{\mu}}^{\mu} \frac{\hat{\mu}}{x(x - \mu_0)} dx \\
&= \Big[\log(x - \mu_0)\Big]_{\hat{\mu}}^{\mu} - \frac{\hat{\mu}}{\mu_0} \Big[2 \log(x - \mu_0) - \log\big(x(x - \mu_0)\big)\Big]_{\hat{\mu}}^{\mu} \\
&= \log\left(\frac{\mu - \mu_0}{\hat{\mu} - \mu_0}\right) + \frac{\hat{\mu}}{\mu_0} \log\left(\frac{\mu(\hat{\mu} - \mu_0)}{\hat{\mu}(\mu - \mu_0)}\right).
\end{aligned}
$$

Replacing in the expression of $\mathcal{H}^*(\mu)$, we get

$$\mathcal{H}^*(\mu) = \kappa^* \hat{\mu}(\mu - \mu_0) \left(\frac{\mu}{\mu - \mu_0}\right)^{\frac{\hat{\mu}}{\mu_0}} \left\{ (\hat{\mu} - \mu_0)^{\frac{\hat{\mu}}{\mu_0} - 1} \hat{\mu}^{-\frac{\hat{\mu}}{\mu_0}} + \int_{\hat{\mu}}^{\mu} (x - \mu_0)^{\frac{\hat{\mu}}{\mu_0} - 1} x^{-\frac{\hat{\mu}}{\mu_0} - 1} dx \right\}.$$

The remaining integral is

$$\int_{\hat{\mu}}^{\mu} (x - \mu_0)^{\frac{\hat{\mu}}{\mu_0} - 1} x^{-\frac{\hat{\mu}}{\mu_0} - 1} dx = \left[\frac{1}{\hat{\mu}} \left(\frac{x - \mu_0}{x}\right)^{\frac{\hat{\mu}}{\mu_0}}\right]_{\hat{\mu}}^{\mu} = \frac{1}{\hat{\mu}} \left(\frac{\mu - \mu_0}{\mu}\right)^{\frac{\hat{\mu}}{\mu_0}} - \frac{1}{\hat{\mu}} \left(\frac{\hat{\mu} - \mu_0}{\hat{\mu}}\right)^{\frac{\hat{\mu}}{\mu_0}}.$$

Finally, we obtain

$$\mathcal{H}^*(\mu) = \kappa^*(\mu - \mu_0) \left\{ 1 + \mu_0(\hat{\mu} - \mu_0)^{\frac{\hat{\mu}}{\mu_0} - 1} \hat{\mu}^{-\frac{\hat{\mu}}{\mu_0}} \left(\frac{\mu}{\mu - \mu_0}\right)^{\frac{\hat{\mu}}{\mu_0}} \right\}. \tag{4}$$

Differentiating, we find

$$H^*(\mu) = \kappa^* \left\{ 1 + \mu_0(\hat{\mu} - \mu_0)^{\frac{\hat{\mu}}{\mu_0} - 1} \hat{\mu}^{-\frac{\hat{\mu}}{\mu_0}} (\mu - \hat{\mu})(\mu - \mu_0)^{-\frac{\hat{\mu}}{\mu_0}} \mu^{\frac{\hat{\mu}}{\mu_0} - 1} \right\}.$$

Hence H^* is continuously differentiable on $[\hat{\mu}, 1]$. We denote its derivative by h^*. Differen-

tiating again, we get

$$h^*(\mu) = \kappa^* \mu_0 (\hat{\mu} - \mu_0)^{\frac{\hat{\mu}}{\mu_0}} \hat{\mu}^{1 - \frac{\hat{\mu}}{\mu_0}} (\mu - \mu_0)^{-\frac{\hat{\mu}}{\mu_0} - 1} \mu^{\frac{\hat{\mu}}{\mu_0} - 2}. \tag{5}$$

Hence $h^*(\mu)$ is strictly positive on $\left[\hat{\mu}, 1\right]$, and H^* is strictly increasing.

To conclude step 1, we only need to show that $H^*(1) \le 1$. By (IDE), we have $H^*(1) = \kappa^* \hat{\mu} + 1 - \hat{\mu}$. Hence, we need to show $\kappa^* \le 1$. Using (4) and the condition $\mathcal{H}^*(1) = 1 - \mu_0$, we have

$$1 - \mu_0 = \mathcal{H}(1) = \kappa^*(1 - \mu_0) \underbrace{\left\{ 1 + \mu_0 (\hat{\mu} - \mu_0)^{\frac{\hat{\mu}}{\mu_0} - 1} \hat{\mu}^{-\frac{\hat{\mu}}{\mu_0} - 2} \left(\frac{1}{1 - \mu_0} \right)^{\frac{\hat{\mu}}{\mu_0}} \right\}}_{\ge 1},$$

which concludes the proof.

Step 3. Suppose that $\mathcal{H}$ is an optimal experiment, that is not less informative than $\mathcal{H}^*$. By Lemma 4, we can as well take $\mathcal{H}$ to be linear since the linear transformation invoked in this lemma is above the original experiment, and therefore more informative. Since $\mathcal{H}$ is optimal, we must have $\mathcal{H}(\mu) = \mathcal{H}^*(\mu) = \kappa^* \mu$, for all $\mu \le \hat{\mu}$. For $\mathcal{H}$ not to be less informative than $\mathcal{H}^*$, there must therefore exist some $\mu \in (\hat{\mu}, 1)$ such that $\mathcal{H}(\mu) > \mathcal{H}^*(\mu)$. Since $\mathcal{H} - \mathcal{H}^*$ is continuous and $\mathcal{H}(1) = \mathcal{H}^*(1)$, we can find the lowest point x above μ at which $\mathcal{H}(x) = \mathcal{H}^*(x)$. Let $\tilde{\mu}$ be this point. Then $\mathcal{H}(x) > \mathcal{H}^*(x)$ for every $x \in [\mu, \tilde{\mu})$. But then, there must exist a subset X of $[\mu, \tilde{\mu}]$ with positive measure, such that $H(x) < H^*(x)$ for every $x \in X$, as otherwise, we would have $\mathcal{H}(\tilde{\mu}) - \mathcal{H}(\mu) = \int_\mu^{\tilde{\mu}} H(\mu) d\mu \ge \int_\mu^{\tilde{\mu}} H^*(\mu) d\mu = \mathcal{H}^*(\tilde{\mu}) - \mathcal{H}^*(\mu)$, a contradiction. Then take $x \in X$. We have $H(x) < H^*(x)$ and $\mathcal{H}(x) > \mathcal{H}^*(x)$. Therefore

$$xH(x) - \frac{x - \hat{\mu}}{x - \mu_0} < xH^*(x) - \frac{x - \hat{\mu}}{x - \mu_0} \mathcal{H}^*(x) = \kappa^* \hat{\mu},$$

and $\mathcal{H}$ must violate (IC$_0'$).

$\square$

Proof of Proposition 7. We have already proved that $\mathcal{H}^*$ is continuously differentiable and ad-

mits a density on $[\hat{\mu}, 1)$, which is given by (5). Differentiating (5), we get

$$h^{*\prime}(\mu) = -\kappa^* \mu_0 (\hat{\mu} - \mu_0)^{\frac{\hat{\mu}}{\mu_0}} \hat{\mu}^{1-\frac{\hat{\mu}}{\mu_0}} (\mu - \mu_0)^{-\frac{\hat{\mu}}{\mu_0}-2} \mu^{\frac{\hat{\mu}}{\mu_0}-3} \big(\mu + \hat{\mu} + 2(\mu - \mu_0) \big) < 0.$$

Note that we can also write

$$h^{*\prime}(\mu) = \frac{h(\mu)}{\mu(\mu - \mu_0)} \Big\{ -\hat{\mu} - \mu - 2(\mu - \mu_0) \Big\}.$$

Differentiating the expressions in Lemma 2, we obtain that the densities of the belief distributions generated by the two types on $[\hat{\mu}, 1)$ are

$$h_G^*(\mu) = \frac{\mu}{\mu_0} h^*(\mu),$$

and

$$h_B^*(\mu) = \frac{1 - \mu}{1 - \mu_0} h(\mu).$$

A quick calculation yields

$$h_G^{*\prime}(\mu) = \frac{h^*(\mu)}{\mu_0(\mu - \mu_0)} \Big\{ -\hat{\mu} - \mu - (\mu - \mu_0) \Big\} < 0,$$

and

$$h_B^{*\prime}(\mu) = \frac{h^*(\mu)}{(1 - \mu_0)(\mu - \mu_0)\mu} \Big\{ -(1 - \mu) \big[\hat{\mu} + \mu + (\mu - \mu_0) \big] - \mu(\mu - \mu_0) \Big\} < 0.$$

To prove first-order stochastic dominance, we can use the expressions in Lemma 2 to get

$$H_G^*(\mu) - H_B^*(\mu) = \frac{1}{\mu_0(1 - \mu_0)} \Big\{ (\mu - \mu_0) H^*(\mu) - \mathcal{H}^*(\mu) \Big\}.$$

We know by Lemma 5 that this expression is negative for $\mu \geq \mu_0$. For $\mu < \mu_0$, we have $H^*(\mu) = \kappa^*$, and $\mathcal{H}^*(\mu) = \kappa^* \mu$, therefore

$$H_G^*(\mu) - H_B^*(\mu) = -\frac{\kappa^*}{1 - \mu_0} < 0.$$

$\square$

Proof of Proposition 8. Pareto efficiency can be seen graphically. Fixing a payoff for the re-

ceiver, that is a value of $\mathcal{F}(\hat{\mu})$, the information structure that maximizes the payoff of the persuader is the one that minimizes the left derivative $F(\hat{\mu})$, while keeping the function $\mathcal{F}$ convex, and under the constraint that $\mathcal{F}(0) = 0$. The only possibility is therefore to make $\mathcal{F}$ linear between $(0, 0)$, and $(\hat{\mu}, \mathcal{F}(\hat{\mu}))$.

For the performance ratio, consider first $\mathcal{H}_{3S}^*$. Recalling that the payoff of the receiver is equal to $\mu_0 - \hat{\mu} + \mathcal{F}(\hat{\mu})$, the performance ratio is

$$\frac{\mu_0 - \hat{\mu} + \kappa_{3S}^* \hat{\mu}}{\mu_0(1 - \hat{\mu})} = \frac{1}{2 - \hat{\mu}}.$$

Interestingly, this ratio is independent of μ_0. It is easy to see that it is bounded below by $1/2$, and that this bound is strict.

Next, the performance ratio of $\mathcal{H}^*$ must by construction be greater than the performance ratio of $\mathcal{H}_{3S}^*$, and hence above $1/2$. To show that this bound is strict, we construct a sequence of pairs $(\mu_0, \hat{\mu})$ such that the corresponding performance ratio approaches $1/2$. The performance ratio of $\mathcal{H}^*$ is given by

$$R(\mu_0, \hat{\mu}) = \frac{\mu_0 - \hat{\mu} + \kappa_{3S}^* \hat{\mu}}{\mu_0(1 - \hat{\mu})} = \frac{\mu_0 - \hat{\mu} + \hat{\mu}\left(1 + \frac{\mu_0}{\hat{\mu} - \mu_0}\left(\frac{\hat{\mu} - \mu_0}{\hat{\mu}(1 - \mu_0)}\right)^{\frac{\hat{\mu}}{\mu_0}}\right)^{-1}}{\mu_0(1 - \hat{\mu})}$$

$$= \frac{1 - \left(\frac{\hat{\mu} - \mu_0}{\hat{\mu}(1 - \mu_0)}\right)^{\frac{\hat{\mu}}{\mu_0}}}{(1 - \hat{\mu})\left(1 + \frac{\mu_0}{\hat{\mu} - \mu_0}\left(\frac{\hat{\mu} - \mu_0}{\hat{\mu}(1 - \mu_0)}\right)^{\frac{\hat{\mu}}{\mu_0}}\right)}.$$

The sequence we consider is defined for $n \geq 2$ by

$$\mu_0^n = \frac{1}{n},$$
$$\hat{\mu}^n = \frac{1}{n} + \frac{1}{n^2}.$$

Hence

$$R\left(\mu_0^n, \hat{\mu}^n\right) = \frac{1}{1 - \hat{\mu}^n} \frac{1 - \left(\frac{n}{(n-1)(n+1)}\right)^{1 + \frac{1}{n}}}{1 + n\left(\frac{n}{(n-1)(n+1)}\right)^{1 + \frac{1}{n}}}$$

As $n \to \infty$, the term $\frac{1}{1-\hat{\mu}^n}$ converges to 1, and the term $\left(\frac{n}{(n-1)(n+1)}\right)^{1+\frac{1}{n}}$ converges to 0. For the remaining term, we can write:

$$n\left(\frac{n}{(n-1)(n+1)}\right)^{1+\frac{1}{n}} = \left(\frac{1}{1+\frac{1}{n}}\right)\left(\frac{1}{1-\frac{1}{n}}\right)^{1+\frac{1}{n}}\left(\frac{1}{1+n}\right)^{\frac{1}{n}}.$$

Since each of the terms in this product converges to 1 as $n \to \infty$, we have

$$\lim_{n\to\infty} R\left(\mu_0^n, \hat{\mu}^n\right) = \frac{1}{2}.$$

$\square$

Proof of Theorem 2. We proceed in three steps.

Step 1: Optimality: Optimality works as in the proof of Theorem 1.

Step 2: $\mathcal{H}_{FI}(\mu) > \mathcal{H}_\lambda^*(\mu) > \mathcal{H}^*(\mu)$. Using the expressions of $\mathcal{H}^*$ and $\mathcal{H}_\lambda^*$, we can write the difference of the two functions for each $\mu \geq \hat{\mu}$ as

$$\mathcal{H}_\lambda^*(\mu) = \mathcal{H}^*(\mu) + \lambda \mu_0 \psi(\mu)\frac{G(\mu)}{G(1)}(B(1) - B(\mu)) \tag{6}$$

where $B(y) \equiv \int_{\hat{\mu}}^y \frac{x-\hat{\mu}}{x(x-\mu_0)\psi(x)}dx$ and $G(y) \equiv \left(1 + \int_{\hat{\mu}}^y \frac{1}{x\psi(x)}dx\right)$ which, because $B(1) - B(\mu) > 0$ and all other terms are positive, implies that $\mathcal{H}_{FI}(\mu) > \mathcal{H}_\lambda^*(\mu)$ on $(0, 1)$.

To see how we can get (6), note that

$$\kappa^* = \frac{1 - \mu_0}{\hat{\mu}\psi(1)\underbrace{\left(1 + \int_{\hat{\mu}}^1 \frac{1}{x\psi(x)}dx\right)}_{G(1)}} = \frac{1 - \mu_0}{\hat{\mu}\psi(1)G(1)} \tag{7}$$

which implies the following expression for $\mathcal{H}^*(\mu)$:

$$\mathcal{H}^*(\mu) = \kappa^*\hat{\mu}\psi(\mu)\left(1 + \int_{\hat{\mu}}^\mu \frac{1}{x\psi(x)}dx\right) = \kappa^*\hat{\mu}\psi(\mu)G(\mu) = (1 - \mu_0)\frac{\psi(\mu)G(\mu)}{\psi(1)G(1)}. \tag{8}$$

47

Note also that

$$\kappa_\lambda^* = \left(\frac{1-\mu_0}{\hat\mu\psi(1)} + \lambda\frac{\mu_0}{\hat\mu} \underbrace{\int_{\hat\mu}^1 \frac{x-\hat\mu}{x(x-\mu_0)\psi(x)}dx}_{B(1)} \right) \left(1 + \int_{\hat\mu}^1 \frac{1}{x\psi(x)}dx \right)^{-1} = \left(\frac{1-\mu_0}{\hat\mu\psi(1)} + \lambda\frac{\mu_0}{\hat\mu}B(1) \right) G(1)^{-1},$$

or, combined with (7):

$$\kappa_\lambda^* = \kappa^* + \lambda\frac{\mu_0}{\hat\mu}\frac{B(1)}{G(1)},$$

which allows us to write:

$$\mathcal{H}_\lambda^*(\mu) = \hat\mu\psi(\mu)\left[\kappa_\lambda^* G(\mu) - \lambda\frac{\mu_0}{\hat\mu}B(\mu) \right]$$

which gives us (6). Now replacing (8) to (6) we obtain:

$$\mathcal{H}_\lambda^*(\mu) = (1-\mu_0)\frac{\psi(\mu)G(\mu)}{\psi(1)G(1)} + \lambda\mu_0\psi(\mu)\frac{G(\mu)}{G(1)}(B(1)-B(\mu)). \tag{9}$$

Finally noting that $\mathcal{H}_{FI}(\mu)$ is a solution to the differential equation when $\lambda = 1 - \mu_0$, (9) implies that $\mathcal{H}_{FI}(\mu) > \mathcal{H}_\lambda^*(\mu)$ when $\lambda < 1 - \mu_0$.

Step 3: $\mathcal{H}_\lambda^* \in \Delta^B$:   After some algebra, we get the following expression of $\mathcal{H}_\lambda^*$ to the right of $\hat\mu$.

$$\mathcal{H}_\lambda^*(\mu) = \kappa_\lambda^*\mu + (\kappa_\lambda^* - \lambda)\mu_0 \left\{ \left(\frac{\mu}{\hat\mu} \right)^{\frac{\hat\mu}{\mu_0}} \left(\frac{\hat\mu - \mu_0}{\mu - \mu_0} \right)^{\frac{\hat\mu}{\mu_0}-1} - 1 \right\}.$$

This implies

$$\kappa_\lambda^* = \frac{1 - \mu_0 + \lambda\mu_0 \left[\hat\mu^{-\frac{\hat\mu}{\mu_0}} \left(\frac{\hat\mu-\mu_0}{1-\mu_0} \right)^{\frac{\hat\mu}{\mu_0}-1} - 1 \right]}{1 - \mu_0 + \mu_0\hat\mu^{-\frac{\hat\mu}{\mu_0}} \left(\frac{\hat\mu-\mu_0}{1-\mu_0} \right)^{\frac{\hat\mu}{\mu_0}-1}} > \lambda.$$

Differentiating, we get

$$H_\lambda^*(\mu) = \kappa_\lambda^* + (\kappa_\lambda^* - \lambda)\mu_0\hat\mu^{-\frac{\hat\mu}{\mu_0}}(\hat\mu-\mu_0)^{\frac{\hat\mu}{\mu_0}-1}(\mu-\hat\mu)\mu^{\frac{\hat\mu}{\mu_0}-1}(\mu-\mu_0)^{-\frac{\hat\mu}{\mu_0}} \geq \kappa_\lambda^*.$$

48

And differentiating again

$$h_\lambda^*(\mu) = (\kappa_\lambda^* - \lambda)\mu_0\hat{\mu}^{-\frac{\hat{\mu}}{\mu_0}}(\hat{\mu} - \mu_0)^{\frac{\hat{\mu}}{\mu_0}}\mu^{\frac{\hat{\mu}}{\mu_0} - 2}(\mu - \mu_0)^{-\frac{\hat{\mu}}{\mu_0} - 1} > 0. \tag{10}$$

Hence, we have convexity. Combined with $\mathcal{H}_{FI}(\mu) > \mathcal{H}_\lambda^*(\mu) > \mathcal{H}^*(\mu)$, this proves that $\mathcal{H}_\lambda^*$ is in Δ^B.

$\square$

Proof of Proposition 9. The proof can be obtained from (10), by proceeding as in the proof of Proposition 7.

$\square$

Proof of Proposition 10. Take $1 - \mu_0 \geq \lambda' > \lambda \geq 0$. Then we can prove $\mathcal{H}_{\lambda'}^*(\mu) > \mathcal{H}_\lambda^*(\mu)$ exactly in the same way as we prove $\mathcal{H}_{FI}(\mu) > \mathcal{H}_\lambda^*(\mu) > \mathcal{H}^*(\mu)$ in the proof of Theorem 2.

$\square$

B Observability and No Limits to Falsification Rates

In this Appendix we explain why removing falsification limits while assuming perfect observability leads to manipulations. Under $\mathcal{H}^*$, choosing $p_B + p_G > 1$ leads the receiver to form beliefs below $\hat{\mu}$ whenever she observes a signal above $\hat{\mu}$. So all signals that would have led to approval under no falsification now lead to rejection. However, the reject signal 0 may now lead to a belief above $\hat{\mu}$. In fact, the optimal falsification rates with $p_B + p_G > 1$ must lead the receiver to form belief $\hat{\mu}$ when she sees signal 0. This optimal falsification strategy is described in the following proposition, and illustrated in Figure 13.

Proposition 12. *Under Assumption 1, but without limits on falsification rates, the optimal falsification strategy under $\mathcal{H}^*$ is to choose $p_G = 1$, and $p_B = \frac{\hat{\mu} - \mu_0}{\hat{\mu}(1 - \mu_0)}$. The persuader gets a payoff of $\mu_0/\hat{\mu}$, whereas the receiver gets a null payoff.*

Proof. Optimality of the proposed falsification strategy among those such that $p_B + p_G > 1$ follows from the arguments just given. Among other falsification strategies, we know that $(0, 0)$ is optimal, by design of $\mathcal{H}^*$. To show that the proposed falsification strategy is optimal among all available ones, we just need to show that the payoff it yields for the persuader, $\mu_0/\hat{\mu}$ is greater
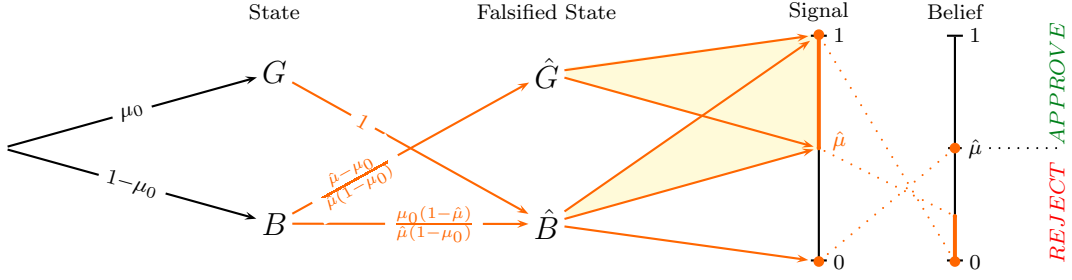
Figure 13: *Manipulating $\mathcal{H}^*$, under perfect observability and no limits on falsification.*

than the payoff the persuader gets under $(0,0)$. The latter is given by $1 - H^*(\hat{\mu}) = 1 - \kappa^*$. Hence we need to show that

$$\kappa^* = \frac{1}{1 + \mu_0(\hat{\mu} - \mu_0)^{\frac{\hat{\mu}}{\mu_0}-1}\hat{\mu}^{-\frac{\hat{\mu}}{\mu_0}}(1-\mu_0)^{-\frac{\hat{\mu}}{\mu_0}}} > \frac{\hat{\mu} - \mu_0}{\hat{\mu}},$$

or, after simplification,

$$1 < \left(\frac{\hat{\mu} - \mu_0}{\hat{\mu}(1-\mu_0)}\right)^{\frac{\hat{\mu}}{\mu_0}},$$

which holds as $\hat{\mu} > \mu_0$. □

Thus, under perfect observability, the persuader can profitably use falsification rates such that $p_B + p_G > 1$ when the test is $\mathcal{H}^*$. But this problem vanishes if we also relax the perfect observability assumption Assumption 1, and instead allow the receiver to learn about manipulations only through the cross-sectional distribution of test results as we do in Section 9.

References

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," .

CHASSANG, S. AND J. ORTNER (2016): "Making Corruption Harder: Asymmetric Information, Collusion and Crime," Working paper.

COHN, J. B., U. RAJAN, AND G. STROBL (2016): "Credit ratings: strategic issuer disclosure and optimal screening," Working paper.

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

CUNNINGHAM, T. AND I. MORENO DE BARREDA (2015): "Equilibrium Persuasion," Working paper.

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 Rothschild-Stiglitz Approach to Bayesian Persuasion," *American Economic Review: Papers and Proceedings*, 106, 597–601.

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*.

HÖRNER, J. AND N. S. LAMBERT (2016): "Motivational ratings," Working paper.

JACKSON, M. O. AND H. F. SONNENSCHEIN (2007): "Overcoming incentive constraints by linking decisions," *Econometrica*, 75, 241–257.

KAMENICA, E. AND M. GENTZKOW (2011): "Bayesian Persuasion," *American Economic Review*, 101, 2590–2615.

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.

MYERSON, R. B. (1991): *Game Theory, Analysis of Conflict*, Harvard University Press.

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.

C   Online Appendix: General Cost Functions

Here, we take back the analysis of optimal design with costly falsification in Section 8 right before introducing the class of linear cost functions. In particular, we consider cost functions $c(p_B)$ defined on I that satisfy Assumption 3.

If (FI) does not hold, the natural intuition is to proceed as in the case without cost. However, the solution of the indifference differential equation with the original cost function may, in general, not be in Δ^B. To circumvent this problem, we work with a modified cost function such that the differential equation always yields a solution in Δ^B, and this solution is optimal for the problem with the original cost function. We obtain this modified cost function recursively. To understand this, it is useful to rewrite the receiver-optimal design program as follows. First, note that Lemma 4 holds with costs, so we can focus on tests $\mathcal{H}$ that are linear to the left of $\hat{\mu}$. Such tests can be parameterized by the slope $\kappa \in \left[1 - \mu_0/\hat{\mu}, 1 - \mu_0\right]$ of the test to the left of $\hat{\mu}$. Then, we have $H(\hat{\mu}) = \mathcal{H}(\hat{\mu})/\hat{\mu} = \kappa$. And, letting Δ_κ^B denote the set of these tests with slope κ to the left of $\hat{\mu}$, we can rewrite the program as

$$\max_{\kappa \in [\kappa^*, 1-\mu_0]} \quad \max_{\mathcal{H} \in \Delta_\kappa^B} \quad \kappa \hat{\mu}$$

$$\text{s.t.} \quad \hat{\mu} c \left(\frac{\mu_0(\mu - \hat{\mu})}{\hat{\mu}(\mu - \mu_0)} \right) \geq \kappa \hat{\mu} + \frac{\mu - \hat{\mu}}{\mu - \mu_0} \mathcal{H}(\mu) - \mu H(\mu), \quad \forall \mu \geq \hat{\mu}. \quad (\text{IC}_0'^c)$$

Note that the optimal no-cost test $\mathcal{H}^*$ satisfies the no-falsification incentive constraint (IC_0^c), so a receiver payoff above $\mathcal{H}^*(\hat{\mu}) = \kappa^* \hat{\mu}$ can be ensured, which is why we limited the range of slopes over which we optimize to $\left[\kappa^*, 1 - \mu_0\right]$.

Next, we show that the cost function can be modified in $(\text{IC}_0'^c)$ without modifying the constraint it puts on all tests in Δ_κ^B. To understand the intuition behind this modification, recall that $(\text{IC}_0'^c)$ simply expresses that the net profit from falsification should be lower than the cost, that is $\Pi(\mu) - \Pi(\hat{\mu}) \leq c \left(\frac{\mu_0(\mu - \hat{\mu})}{\hat{\mu}(\mu - \mu_0)} \right)$. Thus, higher cost helps the receiver achieve better outcomes as they enlarge the set of tests that satisfy the no falsification incentive constraint. However, excessively high costs are unnecessary. To see that consider two falsification levels $p_B < p_B'$ in I that induce thresholds $\mu < \mu'$. Then, we show that the difference in net profits between these two falsification levels, $\Pi(\mu') - \Pi(\mu)$, can be bounded above by $\kappa(p_B' - p_B)$ for all tests in Δ_κ^B. Therefore any cost in excess of $c(p_B) + \kappa(p_B' - p_B)$ at p_B' is superfluous, and can be eliminated without any harm to the receiver.

This intuition leads us to define the modified cost functions on I by

$$\hat{c}_\kappa(x) = \min_{y \in [0,x]} \ c(y) + \kappa(x - y).$$

As stated in the following lemma, working with these modified cost functions is without loss of generality because, due to the intuition outlined above, it leads to an equivalent set of incentive constraints. The proof of the lemma consists in deriving the upper bound that we used in the intuition.

Lemma 6. *Suppose that $\mathcal{H} \in \Delta_\kappa^B$. Then $\mathcal{H}$ satisfies $(\text{IC}_0'^c)$ if and only if it satisfies the same incentive constraint with $\hat{c}_\kappa$, that is*

$$\hat{\mu} \hat{c}_\kappa \left(\frac{\mu_0(\mu - \hat{\mu})}{\hat{\mu}(\mu - \mu_0)} \right) \geq \kappa \hat{\mu} + \frac{\mu - \hat{\mu}}{\mu - \mu_0} \mathcal{H}(\mu) - \mu H(\mu), \quad \forall \mu \geq \hat{\mu}. \quad (\text{IC}_0'^{\hat{c}_\kappa})$$

Proof. Consider two falsification levels $p'_B > p_B$ in I. Let $\mu' > \mu$ be the thresholds they induce in $[\hat{\mu}, 1]$. The difference in net profits between these two levels of falsifications is given by

$$\Pi(\mu') - \Pi(\mu) = \left\{ \frac{\mu' - \hat{\mu}}{\hat{\mu}(\mu' - \mu_0)}\mathcal{H}(\mu') - \frac{\mu'}{\hat{\mu}}H(\mu') \right\} - \left\{ \frac{\mu - \hat{\mu}}{\hat{\mu}(\mu - \mu_0)}\mathcal{H}(\mu) - \frac{\mu}{\hat{\mu}}H(\mu) \right\}.$$

By convexity, $\mathcal{H}(\mu)$ is absolutely continuous, and so is the function $\mu \mapsto \frac{\mu - \hat{\mu}}{\mu - \mu_0}$, therefore we can write the difference between the first terms in each bracket as

$$\frac{\mu' - \hat{\mu}}{\hat{\mu}(\mu' - \mu_0)}\mathcal{H}(\mu') - \frac{\mu - \hat{\mu}}{\hat{\mu}(\mu - \mu_0)}\mathcal{H}(\mu) = \int_\mu^{\mu'} \left\{ \frac{\hat{\mu} - \mu_0}{\hat{\mu}(x - \mu_0)^2}\mathcal{H}(x) + \frac{(x - \hat{\mu})}{\hat{\mu}(x - \mu_0)}H(x) \right\}dx$$

$$= \int_\mu^{\mu'} \frac{\hat{\mu} - \mu_0}{\hat{\mu}(x - \mu_0)^2}\left\{ \mathcal{H}(x) - (x - \mu_0)H(x) \right\}dx,$$

Then, by convexity, we have

$$\frac{\int_\mu^{\mu'} H(x)dx}{\mu' - \mu} = \frac{\mathcal{H}(\mu') - \mathcal{H}(\mu)}{\mu' - \mu} \leq H(\mu')$$

implying

$$-\frac{1}{\hat{\mu}}\int_\mu^{\mu'} H(x)dx \geq -\frac{\mu'}{\hat{\mu}}H(\mu') + \frac{\mu}{\hat{\mu}}H(\mu') \geq -\frac{\mu'}{\hat{\mu}}H(\mu') + \frac{\mu}{\hat{\mu}}H(\mu).$$

Reassembling everything, we have

$$\Pi(\mu') - \Pi(\mu) \leq \int_\mu^{\mu'} \frac{\hat{\mu} - \mu_0}{\hat{\mu}(x - \mu_0)^2}\left\{ \mathcal{H}(x) - (x - \mu_0)H(x) \right\}dx$$

$$\leq \left\{ \mathcal{H}(\hat{\mu}) - (\hat{\mu} - \mu_0)H(\hat{\mu}) \right\}\int_\mu^{\mu'} \frac{\hat{\mu} - \mu_0}{\hat{\mu}(x - \mu_0)^2}dx$$

$$\leq \mu_0\kappa\left\{ \frac{\mu' - \hat{\mu}}{\hat{\mu}(\mu' - \mu_0)} - \frac{\mu - \hat{\mu}}{\hat{\mu}(\mu - \mu_0)} \right\}$$

$$\leq \kappa\left\{ \frac{\mu_0(\mu' - \hat{\mu})}{\hat{\mu}(\mu' - \mu_0)} - \frac{\mu_0(\mu - \hat{\mu})}{\hat{\mu}(\mu - \mu_0)} \right\} = \kappa(p'_B - p_B),$$

where the second inequality is implied by Lemma 5, the third line is due to the linearity of $\mathcal{H}$ to the left of $\hat{\mu}$, which yields $\mathcal{H}(\hat{\mu}) = \hat{\mu}H(\hat{\mu}) = \kappa\hat{\mu}$. $\qquad\square$

The modified cost function satisfies the following technical properties which are crucial in proving that the solution to the differential equation with the modified cost function is in Δ^B.

Lemma 7. *For every $\kappa \in [\kappa^*, 1 - \mu_0]$, the modified cost function $\hat{c}_\kappa(x)$ is well defined, absolutely continuous, nonnegative and nondecreasing on I. It satisfies $\hat{c}_\kappa(0) = 0$, and $\hat{c}_\kappa(x) \leq \min\{\kappa x, c(x)\}$ for every $x \in I$. Furthermore, $\kappa x - \hat{c}_\kappa(x)$ is nondecreasing, and, for $\kappa' > \kappa$, $\hat{c}_{\kappa'}(x) \geq \hat{c}_\kappa(x)$ for every $x \in I$.*

Proof. $\hat{c}_\kappa(\cdot)$ is well defined since the function $y \mapsto c(y) + \kappa(y - x)$ is continuous and therefore admits a minimum on $[0, x]$. $\hat{c}_\kappa(x)$ is nonnegative as the minimum of a nonnegative function.

By definition, $\hat{c}(x) \leq c(0) + \kappa(x - 0) = \kappa x$, and $\hat{c}(x) \leq c(x)$. This implies $\hat{c}(0) = 0$. Let $\hat{y}_\kappa(x) = \arg\min_{y \in [0,x]} c(y) + \kappa(x - y)$. By the maximum theorem, $\hat{y}(\cdot)$ is a nonempty valued correspondence. Consider $x' > x$, and $y' \in \hat{y}_\kappa(x')$. Suppose first that $y' > x$. Then

$$\hat{c}_\kappa(x') = c(y') + \kappa(x' - y') \geq c(y') \geq c(x) \geq \hat{c}_\kappa(x).$$

Suppose, otherwise, that $y' \leq x$. Then

$$\hat{c}_\kappa(x') = c(y') + \kappa(x - y') + \kappa(x' - x) \geq \hat{c}_\kappa(x) + \kappa(x' - x) \geq \hat{c}_\kappa(x).$$

Hence $\hat{c}_\kappa(\cdot)$ is nondecreasing. Next, let $y \in \hat{y}_\kappa(x)$, and note that

$$\hat{c}_\kappa(x') - \hat{c}_\kappa(x) \leq \big[c(y) + \kappa(x' - y)\big] - \big[c(y) - \kappa(x - y)\big] \leq \kappa(x' - x).$$

Therefore, $\hat{c}_\kappa(\cdot)$ is κ-Lipschitz continuous, and in particular absolutely continuous. Furthermore, this implies that $\kappa x - \hat{c}_\kappa(x)$ is nondecreasing.

Next, for $\kappa' > \kappa$, and $y' \in \hat{y}_{\kappa'}(x)$, we have

$$\hat{c}_{\kappa'}(x) = c(y') + \kappa'(x - y') \geq c(y') + \kappa(x - y') \geq \hat{c}_\kappa(x).$$

$\square$

In what follows, to simplify notations, we also write the modified cost functions as a function of the induced threshold

$$\gamma_\kappa(\mu) = \hat{c}_\kappa\left(\frac{\mu_0(\mu - \hat{\mu})}{\hat{\mu}(\mu - \mu_0)}\right).$$

Then, Lemma 6 implies that we can reformulate the optimal design program as

$$\max_{\kappa \in [\kappa^*, 1 - \mu_0]} \quad \max_{\mathcal{H} \in \Delta_\kappa^B} \quad \kappa\hat{\mu}$$

$$\text{s.t.} \quad \hat{\mu}\gamma_\kappa(\mu) \geq \kappa\hat{\mu} + \frac{\mu - \hat{\mu}}{\mu - \mu_0}\mathcal{H}(\mu) - \mu H(\mu), \quad \forall \mu \geq \hat{\mu}.$$

To apply the same idea as in the no-cost case, we would solve the differential equation

$$\hat{\mu}\gamma_\kappa(\mu) = \kappa\hat{\mu} + \frac{\mu - \hat{\mu}}{\mu - \mu_0}\mathcal{H}(\mu) - \mu H(\mu)$$

with initial conditions $H(\hat{\mu}) = \mathcal{H}(\hat{\mu})/\hat{\mu} = \kappa$, and then set κ so that $\mathcal{H}(1) = 1 - \mu_0$. The problem with directly applying this idea is that it leads to a very intractable equation in κ making it difficult to characterize the solution. Furthermore, it is difficult to assess existence or uniqueness of a solution, and even more so, to show that a solution is indeed a test. Therefore, we adopt a different method that characterizes the solution of the optimal design problem recursively as follows.

- $\kappa_0 = 1 - \mu_0$.

- To get κ_{n+1}, we write the following linear differential equation on $[\hat{\mu}, 1]$

$$H(\mu) - \frac{\mu - \hat{\mu}}{\mu(\mu - \mu_0)}\mathcal{H}(\mu) = \frac{\hat{\mu}}{\mu}\big(\kappa - \gamma_{\kappa_n}(\mu)\big),$$

iii

with initial conditions $H(\hat{\mu}) = \mathcal{H}(\hat{\mu})/\hat{\mu} = \kappa$. The solution is then given by

$$\mathcal{H}(\mu) = \hat{\mu}\psi(\mu)\left[\kappa\left(1 + \int_{\hat{\mu}}^{\mu}\frac{1}{x\psi(x)}dx\right) - \int_{\hat{\mu}}^{\mu}\frac{\gamma_{\kappa_n}(x)}{x\psi(x)}dx\right],$$

and we set κ_{n+1} to be the unique value of κ such that $\mathcal{H}(1) = 1 - \mu_0$. That is, we have the following recurrence equation

$$\kappa_{n+1} = \left(\frac{1-\mu_0}{\hat{\mu}\psi(1)} + \int_{\hat{\mu}}^{1}\frac{\gamma_{\kappa_n}(x)}{x\psi(x)}dx\right)\left(1 + \int_{\hat{\mu}}^{1}\frac{1}{x\psi(x)}dx\right)^{-1}. \tag{REC}$$

Finally, we let $\mathcal{H}_n(\mu)$ be the solution to the differential equation with $\kappa = \kappa_{n+1}$.

We show in the next theorem that this sequence always converges, and we can therefore define a limit to the sequence of functions $\mathcal{H}_n$. If the limit of this sequence is a test, that is, if it lies in Δ^B, then it is optimal. However, we need to make another assumption on the cost function to ensure that it is the case.[16]

Assumption 4. *The function $\frac{c(p_B)}{p_B}$ is nonincreasing on I.*

Then, we have the following theorem.

Theorem 4. *If the cost function satisfies* (FI)*, then the optimal test is the fully informative one. Otherwise, the sequence $\{\kappa_n\}$ is decreasing and admits a limit $\kappa_c^* \in (\kappa^*, 1 - \mu_0)$. Then, the function given by*

$$\mathcal{H}_c^*(\mu) = \begin{cases} \kappa_c^*\mu & \text{if } \mu \leq \hat{\mu} \\ \hat{\mu}\psi(\mu)\left[\kappa_c^*\left(1 + \int_{\hat{\mu}}^{\mu}\frac{1}{x\psi(x)}dx\right) - \int_{\hat{\mu}}^{\mu}\frac{\gamma_{\kappa_c^*}(x)}{x\psi(x)}dx\right] & \text{if } \mu \geq \hat{\mu} \end{cases}$$

is an optimal test whenever the cost function satisfies Assumption 4. Furthermore, any other optimal experiment must be linear to the left of $\hat{\mu}$ and less informative than $\mathcal{H}_c^*$. Finally, for all $\mu \in (0,1)$, $\mathcal{H}_{FI}(\mu) > \mathcal{H}_c^*(\mu) > \mathcal{H}^*(\mu)$. If Assumption 4 is not satisfied, then κ_c^* is an upper bound on the modified payoff of the receiver.

Proof. We have already proved the first point. Suppose, therefore that the cost function does not satisfy (FI). We prove the results in the theorem in several steps.

Step 1: convergence of the sequence $\{\kappa_n\}$. To show that the sequence $\{\kappa_n\}$ is decreasing, we proceed by induction. First, note that when the cost function is given by $(1 - \mu_0)p_B$, the fully informative test makes the incentive constraint of the persuader hold with equality at ever $\mu \geq \hat{\mu}$. Therefore, the fully informative test solves the linear differential equation

$$H(\mu) - \frac{\mu - \hat{\mu}}{\mu(\mu - \mu_0)}\mathcal{H}(\mu) = \frac{\hat{\mu}}{\mu}\left(1 - \mu_0 - (1 - \mu_0)\frac{\mu_0(\mu - \hat{\mu})}{\hat{\mu}(\mu - \mu_0)}\right),$$

implying that we have, for all $\mu \geq \hat{\mu}$,

$$(1 - \mu_0)\mu = \mathcal{H}_{FI}(\mu) = \hat{\mu}\psi(\mu)\left[\kappa_0\left(1 + \int_{\hat{\mu}}^{\mu}\frac{1}{x\psi(x)}dx\right) - \int_{\hat{\mu}}^{\mu}\frac{\kappa_0\frac{\mu_0(x-\hat{\mu})}{\hat{\mu}(x-\mu_0)}}{x\psi(x)}dx\right],$$

---

[16]Note that Assumption 4 implies Assumption 3.

and, in particular, at $\mu = 1$

$$\kappa_0 = \left(\frac{1 - \mu_0}{\hat{\mu}\psi(1)} + \int_{\hat{\mu}}^1 \frac{\kappa_0 \frac{\mu_0(x - \hat{\mu})}{\hat{\mu}(x - \mu_0)}}{x\psi(x)}dx\right)\left(1 + \int_{\hat{\mu}}^1 \frac{1}{x\psi(x)}dx\right)^{-1}.$$

By construction, κ_1 is given by

$$\kappa_1 = \left(\frac{1 - \mu_0}{\hat{\mu}\psi(1)} + \int_{\hat{\mu}}^1 \frac{\gamma_{\kappa_0}(x)}{x\psi(x)}dx\right)\left(1 + \int_{\hat{\mu}}^1 \frac{1}{x\psi(x)}dx\right)^{-1}.$$

By [Lemma 7](#), we have

$$\gamma_{\kappa_0}(x) = \hat{c}_{\kappa_0}\left(\frac{\mu_0(x - \hat{\mu})}{\hat{\mu}(x - \mu_0)}\right) \leq \min\left\{\kappa_0 \frac{\mu_0(x - \hat{\mu})}{\hat{\mu}(x - \mu_0)}, c\left(\frac{\mu_0(x - \hat{\mu})}{\hat{\mu}(x - \mu_0)}\right)\right\}.$$

Then, $\gamma_{\kappa_0}(x) \leq \kappa_0 \frac{\mu_0(x-\hat{\mu})}{\hat{\mu}(x-\mu_0)}$ for all $x \in [\hat{\mu}, 1]$, and because $c(\cdot)$ does not satisfy (FI), and is continuous, there exists an open interval over which the inequality is strict. Therefore, we must have $\kappa_1 < \kappa_0 = 1 - \mu_0$.

Next, suppose that for $n \geq 1$, we have $\kappa_n \leq \kappa_{n-1}$. Then, [Lemma 7](#) implies that we have $\gamma_{\kappa_n}(x) \leq \gamma_{\kappa_{n-1}}(x)$, for all $x \in [\hat{\mu}, 1]$, and therefore, by (REC), $\kappa_{n+1} \leq \kappa_n$.

Next, note the definition of κ^* implies that, for all $n \geq 0$, $\kappa_n > \kappa^*$. $\{\kappa_n\}$ is therefore a decreasing sequence bounded from below, hence it must converge to a limit $\kappa_c^* \in [\kappa^*, 1 - \mu_0)$. Furthermore, κ_c^* must be a fixed point of the recurrence equation (REC). Therefore

$$\kappa_c^* = \left(\frac{1 - \mu_0}{\hat{\mu}\psi(1)} + \int_{\hat{\mu}}^1 \frac{\gamma_{\kappa_c^*}(x)}{x\psi(x)}dx\right)\left(1 + \int_{\hat{\mu}}^1 \frac{1}{x\psi(x)}dx\right)^{-1},$$

and, since $\kappa_c^* > 0$, $\gamma_{\kappa_c^*}(x) > 0$, for all $x > \hat{\mu}$, implying that $\kappa_c^* > \kappa^*$.

Step 2: $\mathcal{H}_{FI}(\mu) > \mathcal{H}_c^*(\mu) > \mathcal{H}^*(\mu)$. Using the expressions of $\mathcal{H}^*$ and $\mathcal{H}_c^*$, we can write the difference of the two functions for each $\mu \geq \hat{\mu}$ as

$$\mathcal{H}_c^*(\mu) - \mathcal{H}^*(\mu) = \frac{\hat{\mu}\int_{\hat{\mu}}^1 \frac{\gamma_{\kappa_c^*}(x)}{x\psi(x)}dx}{1 + \int_{\hat{\mu}}^1 \frac{1}{x\psi(x)}dx}\psi(\mu)\int_{\hat{\mu}}^\mu \frac{\gamma_{\kappa_c^*}(x)}{x\psi(x)}dx$$

$$\times \underbrace{\left\{\frac{1 + \int_{\hat{\mu}}^\mu \frac{1}{x\psi(x)}dx}{\int_{\hat{\mu}}^\mu \frac{\gamma_{\kappa_c^*}(x)}{x\psi(x)}dx} - \frac{1 + \int_{\hat{\mu}}^1 \frac{1}{x\psi(x)}dx}{\int_{\hat{\mu}}^1 \frac{\gamma_{\kappa_c^*}(x)}{x\psi(x)}dx}\right\}}_{\equiv \Delta(\mu)}, \tag{11}$$

where the second equality is from the proof of [Theorem 1](#).

Note that we have $\Delta(1) = 0$. To assess the sign of this term, we compute its derivative

$$\Delta'(\mu) = \left(\frac{1}{\int_{\hat{\mu}}^\mu \frac{\gamma_{\kappa_c^*}(x)}{x\psi(x)}dx}\right)^2 \frac{1}{\mu\psi(\mu)}\left\{\int_{\hat{\mu}}^\mu \frac{\gamma_{\kappa_c^*}(x)}{x\psi(x)}dx - \gamma_{\kappa_c^*}(\mu)\left(1 + \int_{\hat{\mu}}^\mu \frac{1}{x\psi(x)}\right)\right\}.$$

Since $\gamma_{\kappa_c^*}(\cdot)$ is nondecreasing, we have $\int_{\hat{\mu}}^{\mu} \frac{\gamma_{\kappa_c^*}(x)}{x\psi(x)} dx \leq \gamma_{\kappa_c^*}(\mu) \int_{\hat{\mu}}^{\mu} \frac{1}{x\psi(x)} dx$, and therefore, $\Delta'(\mu) < 0$ on $(\hat{\mu}, 1]$, implying that $\mathcal{H}_c^*(\mu) > \mathcal{H}^*(\mu)$ on $[\hat{\mu}, 1)$, which easily extends to $(0, \hat{\mu}]$ by linearity of both functions on this interval and continuity at $\hat{\mu}$.

Next, note that the fully informative test $\mathcal{H}_{FI}$ is the solution of the differential equation with cost when the cost function is given by $\gamma_{FI}(\mu) = (1 - \mu_0) \frac{\mu_0(\mu - \hat{\mu})}{\hat{\mu}(\mu - \mu_0)}$. Hence, we can write the following version of (11),

$$\mathcal{H}_{FI}(\mu) - \mathcal{H}^*(\mu) = \frac{\hat{\mu} \int_{\hat{\mu}}^{1} \frac{\gamma_{FI}(x)}{x\psi(x)} dx}{1 + \int_{\hat{\mu}}^{1} \frac{1}{x\psi(x)} dx} \psi(\mu) \int_{\hat{\mu}}^{\mu} \frac{\gamma_{FI}(x)}{x\psi(x)} dx$$
$$\times \left\{ \frac{1 + \int_{\hat{\mu}}^{\mu} \frac{1}{x\psi(x)} dx}{\int_{\hat{\mu}}^{\mu} \frac{\gamma_{FI}(x)}{x\psi(x)} dx} - \frac{1 + \int_{\hat{\mu}}^{1} \frac{1}{x\psi(x)} dx}{\int_{\hat{\mu}}^{1} \frac{\gamma_{FI}(x)}{x\psi(x)} dx} \right\}. \tag{12}$$

Subtracting (11) from (12)

$$\mathcal{H}_{FI}(\mu) - \mathcal{H}_c^*(\mu) = \frac{\hat{\mu} \int_{\hat{\mu}}^{1} \frac{\delta(x)}{x\psi(x)} dx}{1 + \int_{\hat{\mu}}^{1} \frac{1}{x\psi(x)} dx} \psi(\mu) \int_{\hat{\mu}}^{\mu} \frac{\delta(x)}{x\psi(x)} dx$$
$$\times \underbrace{\left\{ \frac{1 + \int_{\hat{\mu}}^{\mu} \frac{1}{x\psi(x)} dx}{\int_{\hat{\mu}}^{\mu} \frac{\delta(x)}{x\psi(x)} dx} - \frac{1 + \int_{\hat{\mu}}^{1} \frac{1}{x\psi(x)} dx}{\int_{\hat{\mu}}^{1} \frac{\delta(x)}{x\psi(x)} dx} \right\}}_{\equiv \tilde{\Delta}(\mu)}, \tag{13}$$

where $\delta(x) = \gamma_{FI}(x) - \gamma_{\kappa_c^*}(x)$ is bounded below by 0, above by $\gamma_{FI}(x)$. Lemma 7 implies that $\delta(x)$ is non decreasing in x. Therefore, applying the same argument as for Δ, we can show that $\mathcal{H}_{FI}(\mu) > \mathcal{H}_c^*(\mu)$ on $(0, 1)$.

Step 3: $\mathcal{H}_c^* \in \Delta^B$: Next, we show that $\mathcal{H}_c^* \in \Delta^B$. Given that we already have $\mathcal{H}_{FI}(\mu) > \mathcal{H}_c^*(\mu) > \mathcal{H}^*(\mu)$, it is sufficient to show that $\mathcal{H}_c^*$ is convex to ensure that it is in Δ^B. Using the same computations as in the case without cost, we can write

$$\mathcal{H}_c^*(\mu) = \kappa_c^*(\mu - \mu_0) \left\{ 1 + \mu_0(\hat{\mu} - \mu_0)^{\frac{\hat{\mu}}{\mu_0} - 1} \hat{\mu}^{-\frac{\hat{\mu}}{\mu_0}} \left(\frac{\mu}{\mu - \mu_0} \right)^{\frac{\hat{\mu}}{\mu_0}} \right\}$$
$$- (\mu - \mu_0)^{1 - \frac{\hat{\mu}}{\mu_0}} \mu^{\frac{\hat{\mu}}{\mu_0}} \hat{\mu} \int_{\hat{\mu}}^{\mu} \gamma_{\kappa_c^*}(x)(x - \mu_0)^{\frac{\hat{\mu}}{\mu_0} - 1} x^{-\frac{\hat{\mu}}{\mu_0} - 1} dx.$$

We introduce the function

$$\varphi_\kappa(\mu) = \kappa p_B - \hat{c}_\kappa(p_B) = \kappa \frac{\mu_0(\mu - \hat{\mu})}{\hat{\mu}(\mu - \mu_0)} - \gamma_\kappa(\mu).$$

By Lemma 7, this function is nonnegative and nondecreasing in p_B, and hence in μ. Then, we can rewrite $\mathcal{H}_c^*$ as follows

$$\mathcal{H}_c^*(\mu) = \kappa_c^*\mu + (\mu - \mu_0)^{1-\frac{\hat{\mu}}{\mu_0}}\mu^{\frac{\hat{\mu}}{\mu_0}}\hat{\mu}\Bigg\{ \frac{\kappa\mu_0}{\hat{\mu}}\underbrace{\left(\hat{\mu}^{-\frac{\hat{\mu}}{\mu_0}}(\hat{\mu}-\mu_0)^{\frac{\hat{\mu}}{\mu_0}-1} - \mu^{-\frac{\hat{\mu}}{\mu_0}}(\mu-\mu_0)^{\frac{\hat{\mu}}{\mu_0}-1}\right)}_{=\int_{\hat{\mu}}^{\mu}(x-\hat{\mu})(x-\mu_0)^{\frac{\hat{\mu}}{\mu_0}-2}x^{-\frac{\hat{\mu}}{\mu_0}-1}dx}$$
$$- \int_{\hat{\mu}}^{\mu}\gamma_{\kappa_c^*}(x)(x-\mu_0)^{\frac{\hat{\mu}}{\mu_0}-1}x^{-\frac{\hat{\mu}}{\mu_0}-1}dx.\Bigg\}$$

Therefore

$$\mathcal{H}_c^*(\mu) = \kappa_c^*\mu + (\mu - \mu_0)^{1-\frac{\hat{\mu}}{\mu_0}}\mu^{\frac{\hat{\mu}}{\mu_0}}\hat{\mu}\int_{\hat{\mu}}^{\mu}\varphi_{\kappa_c^*}(x)(x-\mu_0)^{\frac{\hat{\mu}}{\mu_0}-1}x^{-\frac{\hat{\mu}}{\mu_0}-1}dx. \tag{14}$$

Differentiating, we get

$$H_c^*(\mu) = \kappa_c^* + \hat{\mu}\left\{ \frac{\varphi_{\kappa_c^*}(\mu)}{\mu} + (\mu - \hat{\mu})(\mu - \mu_0)^{-\frac{\hat{\mu}}{\mu_0}}\mu^{\frac{\hat{\mu}}{\mu_0}-1}\int_{\hat{\mu}}^{\mu}\varphi_{\kappa_c^*}(x)(x-\mu_0)^{\frac{\hat{\mu}}{\mu_0}-1}x^{-\frac{\hat{\mu}}{\mu_0}-1}dx\right\}. \tag{15}$$

Note that this implies that $H_c^*(\mu) \geq \kappa_c^*$ for all $\mu \geq \hat{\mu}$. Next, note that, by definition, the function $\mathcal{H}_c^*$ solves the differential equation

$$\frac{\mu - \hat{\mu}}{\mu - \mu_0}\mathcal{H}_c^*(\mu) - \mu H_c^*(\mu) + \kappa_c^*\hat{\mu} = \hat{\mu}\gamma_{\kappa_c^*}(\mu),$$

which we can also write

$$\frac{\mu - \hat{\mu}}{\mu - \mu_0}\left(\mathcal{H}_c^*(\mu) - (\mu - \mu_0)H_c^*(\mu)\right) - \hat{\mu}\left(H_c^*(\mu) - \kappa_c^*\right) = \hat{\mu}\gamma_{\kappa_c^*}(\mu).$$

Differentiating this equation, we obtain

$$\mu h_c^*(\mu) = \frac{\hat{\mu} - \mu_0}{(\mu - \mu_0)^2}\left(\mathcal{H}_c^*(\mu) - (\mu - \mu_0)H_c^*(\mu)\right) - \hat{\mu}\gamma'_{\kappa_c^*}(\mu)$$
$$= \frac{\hat{\mu} - \mu_0}{(\mu - \mu_0)(\mu - \hat{\mu})}\left\{ H_c^*(\mu) - \kappa + \gamma_{\kappa_c^*}(\mu) - \frac{(\mu - \mu_0)(\mu - \hat{\mu})}{\hat{\mu} - \mu_0}\gamma'_{\kappa_c^*}(\mu)\right\}$$
$$= \frac{\hat{\mu} - \mu_0}{(\mu - \mu_0)(\mu - \hat{\mu})}\left\{ H_c^*(\mu) - \kappa_c^* + \hat{c}_{\kappa_c^*}(p_B) - p_B\hat{c}'_{\kappa_c^*}(p_B)\right\}$$

We have already proved that $H_c^*(\mu) - \kappa_c^* \geq 0$, and it is easy to see that Assumption 4 implies that $\hat{c}_{\kappa_c^*}(p_B)/p_B$ is nonincreasing, and therefore $\hat{c}_{\kappa_c^*}(p_B) - p_B\hat{c}'_{\kappa_c^*}(p_B) \geq 0$.

Step 4: Optimality of $\mathcal{H}_c^*$: Let $\mathcal{H} \in \Delta_B$ be a test with $\mathcal{H}(\hat{\mu}) = \hat{\mu}\kappa'$, and $\kappa' > \kappa_c^*$ that satisfies the no-falsification incentive constraint. By Lemma 4, we can take this test to be linear to the left of $\hat{\mu}$, that is $\mathcal{H} \in \Delta_\kappa^B$. Then $\mathcal{H}$ satisfies, for every $\mu \geq \hat{\mu}$,

$$\hat{\mu}\gamma_{\kappa'}(\mu) \geq \kappa'\hat{\mu} + \frac{\mu - \hat{\mu}}{\mu - \mu_0}\mathcal{H}(\mu) - \mu H(\mu).$$

Since $\kappa' > \kappa_c^*$, there must exist some $n \geq 0$ such that $\kappa_n \geq \kappa' > \kappa_{n+1}$. Then, by [Lemma 7](), $\gamma_{\kappa_n}(\mu) \geq \gamma_{\kappa'}(\mu)$, implying that the no-falsification incentive constraint must hold with γ_{κ_n} as well, that is, for every $\mu \geq \hat{\mu}$,

$$\hat{\mu}\gamma_{\kappa_n}(\mu) \geq \kappa'\hat{\mu} + \frac{\mu - \hat{\mu}}{\mu - \mu_0}\mathcal{H}(\mu) - \mu H(\mu).$$

Next consider the function $\mathcal{H}_n(\mu)$, which, by definition, satisfies $\mathcal{H}_n(\hat{\mu}) = \hat{\mu}\kappa_{n+1}$, and $\mathcal{H}_n(1) = 1 - \mu_0$, and, for every $\mu \geq \hat{\mu}$,

$$\hat{\mu}\gamma_{\kappa_n}(\mu) = \kappa_{n+1}\hat{\mu} + \frac{\mu - \hat{\mu}}{\mu - \mu_0}\mathcal{H}_n(\mu) - \mu H_n(\mu).$$

Since $\mathcal{H}(\hat{\mu}) > \mathcal{H}_n(\hat{\mu})$, and $\mathcal{H}(1) = \mathcal{H}_n(1) = 1 - \mu_0$, there exists some $\tilde{\mu} \in (\hat{\mu}, 1]$, such that $\mathcal{H}(\tilde{\mu}) = \mathcal{H}_n(\tilde{\mu})$, and $\mathcal{H}(\mu) > \mathcal{H}_n(\mu)$ for $\mu \in [\hat{\mu}, \tilde{\mu})$. But then, we must have $H(\tilde{\mu}) \leq H_n(\tilde{\mu})$. Therefore

$$\hat{\mu}\gamma_{\kappa_n}(\tilde{\mu}) \geq \kappa'\hat{\mu} + \frac{\tilde{\mu} - \hat{\mu}}{\tilde{\mu} - \mu_0}\mathcal{H}(\tilde{\mu}) - \tilde{\mu}H(\tilde{\mu})$$

$$> \kappa_{n+1}\hat{\mu} + \frac{\tilde{\mu} - \hat{\mu}}{\tilde{\mu} - \mu_0}\mathcal{H}_n(\tilde{\mu}) - \tilde{\mu}H_n(\tilde{\mu}) = \hat{\mu}\gamma_{\kappa_n}(\tilde{\mu}),$$

a contradiction. □

Thus, our recursive approach delivers the optimal test whenever [Assumption 4]() is satisfied. When [Assumption 4]() is not satisfied, the recursive approach still delivers a limit function $\mathcal{H}_c^*$. However, we cannot ensure that this function is convex, and therefore corresponds to a test. But it is still true that any optimal test $\mathcal{H}(\mu)$ must lie below $\mathcal{H}_c^*(\mu)$, and therefore the modified payoff of the receiver is bounded above by $\mathcal{H}_c^*(\hat{\mu}) = \kappa_c^*\hat{\mu}$. Furthermore, for any cost function, if $\mathcal{H}_c^*$ happens to be convex so that it is a test, then it is an optimal test.