Supplement to
Test Design under Falsification

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

January 19, 2022

Contents

[*] Sciences Po, CEPR – e-mail: eduardo.perez@sciencespo.fr

[†] UT Austin, UCL, CEPR – e-mail: vskreta@gmail.com

S1 Covert falsification: omitted results, proofs

S1.1 Preliminary results

Lemma S1.1 (Recommendation principle)**.** *Let (ϕ, α) be an equilibrium under test τ. Then, (ϕ, δ^A) is an equilibrium under the test $\alpha\tau$ with signal space $X = A$. Furthermore, the outcomes of both equilibria are identical.*

Proof of Lemma S1.1. The last point is immediate since $\delta^A(\alpha\tau)\phi = \alpha\tau\phi$. Then, first note

$$\mu(\mathfrak{a}|\alpha\tau, \phi) = \frac{\int_{X\times S} \alpha(\mathfrak{a}|x)\mu(x|\tau, \phi) d\tau\phi\pi}{\alpha\tau\phi\pi(\{\mathfrak{a}\} \times S)} \geq 0$$

and

$$\mu(\mathfrak{r}|\alpha\tau, \phi) = \frac{\int_{X\times S} \alpha(\mathfrak{r}|x)\mu(x|\tau, \phi) d\tau\phi\pi}{\alpha\tau\phi\pi(\{\mathfrak{r}\} \times S)} \leq 0$$

because α is a best-response to ϕ under τ. Therefore, δ^A is a best-response to ϕ under $\alpha\tau$. Next note

$$\Pi\big(\delta^A(\alpha\tau)\phi'\big) - C(\phi') = \Pi\big(\alpha\tau\phi'\big) - C(\phi') \leq \Pi\big(\alpha\tau\phi\big) - C(\phi) = \Pi\big(\delta^A(\alpha\tau)\phi\big) - C(\phi),$$

so ϕ is a best-response to δ^A under $\alpha\tau$. □

Lemma S1.2 (Equivalence of ex-ante and interim falsification)**.** *An agent's falsification strategy satisfies* (EF') *if and only if it satisfies* (IEF) *for almost every s.*

Proof of Lemma S1.2. Sufficiency is immediate, so we prove necessity. Suppose a set of states S' with $\pi(S') > 0$ exists such that for every $s \in S'$, $\phi\big(\text{argmax}_t\ \tau(t) - \gamma c(t|s)|s\big) < 1$. Then, for every $s \in S'$, a state $t(s)$ exists such that $\tau(t(s)) - \gamma c(t(s)|s) > \int_{\text{supp}\,\phi(s)}\{\tau(t) - \gamma c(t|s)\} d\phi(t|s)$. Suppose the agent deviates to

$$\tilde{\phi} = \begin{cases} \delta_{t(s)} \text{ for } s \in S' \\ \phi(s) \text{ otherwise .} \end{cases}$$

Because this deviation is covert, the receiver's behavior stays the same. Given that S' has strictly positive measure, the agent's payoff strictly increases:

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

so that ϕ cannot satisfy (EF'). □

S1.2 Comparative statics

Proof of Proposition 2. In this proof, we let $\omega_{\gamma,p,\hat{s}}$ denote the outcome function $\omega_{p,\hat{s}}$ to highlight its dependence on the cost parameter γ. We consider an increase from γ to $\gamma' > \gamma$ that remains in the same cost region.

In the low-cost region, we have $\tau^*_{\gamma'}(s) > \tau^*_\gamma(s)$ for all $s > s_0$, whereas all other states are rejected with certainty under both values. This change is strictly beneficial for the agent. For the receiver, note first that $\omega^*_{\gamma'}$ must, by definition, deliver a higher payoff than $\omega_{\gamma',p^*_\gamma,\overline{s}}$. Next, note $\omega_{\gamma',p^*_\gamma,\overline{s}}$ assigns the same approval probabilities as ω^*_γ to all compliant states but lower (strictly for a positive mass) approval probabilities to noncompliant states, and therefore gives the receiver a strictly higher payoff.

In the intermediate-cost region, $\omega^*_{\gamma'}$ assigns the same approval probability as ω^*_γ to compliant states but a lower one to all noncompliant states (strictly for a positive mass of them). Therefore, the receiver's payoff increases. We also have $\tau^*_{\gamma'}(s) \leq \tau^*_\gamma(s)$ with a strict inequality for a positive mass of states. The agent is therefore worse off under γ'.

In the high-cost region, the receiver gets her first-best payoff, which is independent of γ. For the agent, we have $\tau^*_{\gamma'}(s) \geq \tau^*_\gamma(s)$ with a strict inequality for a positive mass of states, so she is better off under γ'. □

S1.3 Characterization of optimal tests

We provide a characterization of optimal tests that shows, in particular, productive falsification is needed for optimality. To simplify exposition, we exclude from the proposition the case in which $\mu_\pi = 0$. Indeed, this case has multiple tests from the class $\tau_{p,\hat{s}}$ that are optimal in the low-cost region $\gamma c(\overline{s}| - \underline{s}) \leq 1$. Then, any test $\tau_{p,\overline{s}}$ with $p \in [p^*_\gamma, 1]$ is optimal. When $\mu_\pi \neq 0$, the optimal test within our class is unique.

Proposition S1.1. *Suppose $\mu_\pi \neq 0$. A test τ is such that, for some falsification strategy ϕ, (τ, ϕ) solves ($\mathcal{P}$), and $U(\tau\phi, \phi) \geq U^*_\gamma$ if and only if:*

(i) For a.e. $s \in \big[-\underline{s}, 0\big) \cup \big(\hat{s}^*_\gamma, \overline{s}\big]$, $\tau(s) = \tau^*_\gamma(s)$.

(ii) For every $s \in \big[-\underline{s}, \overline{s}\big)$, $\tau(s) \leq \tau^*_\gamma(s)$.

(iii) $\tau\left(\hat{s}^*_\gamma\right) = p^*_\gamma$.

Furthermore, it is then the case that for a.e. s, $\phi(s) = \phi^*_\gamma(s)$.

Proof.

Sufficiency. Suppose τ satisfies (i)-(iii). We show ϕ^*_γ satisfies (EF') under τ. It is then easy to see that $\tau\phi^*_\gamma(s) = \omega^*_\gamma(s)$ a.e., which implies the result.

Let $T = \{s \in S : \tau(s) \neq \tau^*_\gamma(s)\}$. First, consider $s \in S \setminus T$. Then, ϕ^*_γ satisfies (IEF) at s under τ since, first, s can still not falsify and get the same payoff as under τ^*_γ because $s \notin T$; and second, by (ii), no other falsification targets yield higher payoff under τ than under τ^*_γ.

Next, consider $s \in T \cap \left[0, \hat{s}^*_\gamma\right)$. Then, ϕ^*_γ satisfies (IEF) at s under τ because, first, s can still falsify to the standard and get the same payoff as under τ^*_γ by (iii); and second, by (ii), no falsification targets that a higher payoff under τ than under τ^*_γ.

By (i), $T \cap \left(\left[-\underline{s}, 0\right) \cup \left(\hat{s}^*_\gamma, \overline{s}\right]\right)$ has measure 0. Therefore, we have shown ϕ^*_γ satisfies (IEF) almost everywhere under τ, and we can conclude by Lemma S1.2.

Necessity. Let $\omega = \tau\phi$ be an equilibrium outcome that satisfies $V(\omega) = V^*_\gamma$ and $U(\omega, \phi) \geq U^*_\gamma$. To prove necessity, we apply to τ the construction in the first step of the proof of Theorem 1. We refer to this construction as $\mathcal{C}$. Because we can easily see that when optimizing within our class of tests, the solution is always unique if $\mu_\pi \neq 0$, applying $\mathcal{C}$ to τ must yield τ^*_γ.

Let $T_i = \{s \in \left[-\underline{s}, 0\right) \cup \left(\hat{s}^*_\gamma, \overline{s}\right] : \tau(s) \neq \tau^*_\gamma(s)\}$ and $T_{ii} = \{s \in S : \tau(s) > \tau^*_\gamma(s)\}$. We show T_i must have measure 0, $T_{ii} = \emptyset$, so (i) and (ii) must hold, and τ must satisfy (iii).

For every $s \in T_{ii}$, we must have $\tau(s) \leq p^*_\gamma$; otherwise, we could apply $\mathcal{C}$ to get a test $\tau_{p,\hat{s}}$ in our class with $p > p^*_\gamma$, that is $\tau_{p,\hat{s}} \neq \tau^*_\gamma$, a contradiction. This argument implies $T_{ii} \subseteq \left[-\underline{s}, \hat{s}^*_\gamma\right)$.

Suppose $t \in T_{ii}$ exists such that $t < 0$. Then, note that for every s, we must have

$$\omega(s) \geq \sup_{\tilde{t}} \; \tau(\tilde{t}) - \gamma c(\tilde{t}|s) \geq \tau(t) - \gamma c(t|s),$$

and because $\tau(t) - \gamma c(t|s)$ is continuous in s, and $\omega(t) \geq \tau(t) > \tau^*_\gamma(s) = \omega^*_\gamma(s)$, a neighborhood N of t exists such that for every $s \in N$, $\omega(s) > \omega^*_\gamma$ and $s < 0$. Hence, states in N are noncompliant and get a strictly higher approval probability under ω than under ω^*_γ. Because applying $\mathcal{C}$ to τ must yield the outcome ω^*_γ, this argument shows $V(\omega) < V^*_\gamma$, a contradiction. Hence, $T_{ii} \subseteq \left[0, \hat{s}^*_\gamma\right)$.

Next, fix some sufficiently small $\varepsilon > 0$. We show $T_{ii} \subseteq \big(\hat{s}^*_\gamma - \varepsilon, \hat{s}^*_\gamma\big)$. For illustration, let $\tilde{T} = T_{ii} \cap \big[0, \hat{s}^*_\gamma - \varepsilon\big]$. Then, we must have $\tilde{p} = \sup_{t \in \tilde{T}} \tau(t) < p^*_\gamma$, because $\mathcal{C}$ would otherwise yield a test $\tau_{p,\hat{s}} \neq \tau^*_\gamma$. Take $\tilde{s} \in \tilde{T}$ and let $N \subseteq \big[0, \hat{s}^*_\gamma - \varepsilon\big]$ be a small neighborhood of $\tilde{s}$. Then, the payoff of falsifying to some target $t > \hat{s}^*_\gamma - \varepsilon$ for any $s \in N$ is bounded above by $p^*_\gamma - c\big(\hat{s}^*_\gamma - \varepsilon | s\big)$, which is itself strictly lower than $\tau\big(\tilde{s}\big)$, when both N and ε are sufficiently small. Therefore, states in N must be falsifying to targets below $\hat{s}^*_\gamma - \varepsilon$, implying their equilibrium approval probability $\omega(s)$ is at most $\tilde{p} < p^*_\gamma$. But then, because these states are positive, $V(\omega) < V^*_\gamma$, a contradiction.

Overall, we have that, for every $\varepsilon > 0$ sufficiently small, $T_{ii} \subseteq \big(\hat{s}^*_\gamma - \varepsilon, \hat{s}^*_\gamma\big)$. This finding implies $T_{ii} = \emptyset$, so (ii) must hold.

Because $\mathcal{C}$ must yield τ^*_γ, a sequence s_n of states must exist that converges to $\hat{s}^*_\gamma$ and such that $\tau(s_n)$ converges to p^*_γ. Because T_{ii} is empty, falsifying to s_n gives to any $s \in [0, \hat{s}^*_\gamma\big)$ a payoff that converges to $p^*_\gamma - c\big(\hat{s}^*_\gamma | s\big)$, which is an upper bound on what s can achieve by falsifying under τ. To ensure an optimal falsification strategy exists for s that does as well as this sequence, it must be the case that $\tau\big(\hat{s}^*_\gamma\big) = p^*_\gamma$, so (iii) must hold.

Next suppose an open interval I of noncompliant states exists such that $I \subseteq T_i$. Then, for every state $s \in I$, $\tau(s) < \tau^*_\gamma(s)$, so falsifying to $\hat{s}^*_\gamma$ yields a strictly higher payoff than truth-telling, and because (ii) holds, this falsification strategy is optimal for s. Therefore, under ω, all states in $s \in I$ are approved with probability $\omega(s) = p^*_\gamma > \omega^*_\gamma(s)$, implying $V(\tau\phi) < V^*_\gamma$, a contradiction. Hence, $T_i \cap \big[-\underline{s}, 0\big)$ has measure 0.

Next, suppose an open interval $I \subseteq T_i \cap \big(\hat{s}^*_\gamma, \overline{s}\big]$ of states exists. For every $s \in I$, $\tau(s) < p^*_\gamma$, and the only way that ω is a receiver-optimal outcome is if each of these states falsifies to some other state t with $\tau(t) = p^*_\gamma$. However, this argument implies falsifying outside of I, and falsifying all states in I to states outside of I has a strictly positive falsification cost. This observation implies the overall falsification cost under ϕ exceeds the overall falsification cost, whereas all states obtain the same outcome as under ω^*_γ. Therefore, $U(\omega, \phi) < U^*_\gamma$, a contradiction. Therefore, $T_i \cap \big(\hat{s}^*_\gamma, \overline{s}\big]$ must also have measure 0, so (i) must hold. □

S1.4 Upward-only falsification

We show the test of Theorem 1 remains optimal if, instead of (CTT), we assume falsification is *upward-only*; that is, the agent can only falsify state s as states in

$[s, \overline{s}]$.

Theorem S1.1. *Suppose the cost function satisfies* (UTI) *and that falsification is upward-only. Then,* $(\tau^*_\gamma, \phi^*_\gamma)$ *solves* ($\mathcal{P}$)*.*

Proof. Optimization within our class of tests is exactly as in the proof of Theorem 1. We show how to adapt the first part of the proof that shows optimality of our class of tests.

Suppose (τ, ϕ) satisfies (IEF). Let $p = \sup_{s \in S^+} \tau(s)$, which exists because $\tau(\cdot)$ is bounded. For every $\varepsilon > 0$, let $S^+(\varepsilon) = \big\{ s \in S^+ : \tau(s) \geq p - \varepsilon \big\}$, and let $\bar{S}^+(\varepsilon)$ be the closure of $S^+(\varepsilon)$. By definition of p, each $S^+(\varepsilon)$, and hence each $\bar{S}^+(\varepsilon)$, is nonempty. Furthermore, $\bar{S}^+(\varepsilon)$ is clearly nonincreasing in ε for the inclusion order. Therefore, by Cantor's intersection theorem, $\bar{S}^+ = \bigcap_{\varepsilon > 0} \bar{S}^+(\varepsilon)$ is a nonempty compact subset of S^+.

If some $s \in S^+$ exists such that $\gamma c(s|0) \geq p$, we can set $\hat{s} \in S^+$ to be the unique state such that $\gamma c(\hat{s}|0) = p$. Then, under $\omega_{p,\hat{s}}$, every compliant state is approved with probability p, whereas every noncompliant state is rejected with certainty, making the receiver at least as well off as under $\tau\phi$.

Otherwise, we let $\hat{s}$ be the minimal element of $\bar{S}^+$. Then, under $\omega_{p,\hat{s}}$, every compliant state is approved with probability p. Because falsification is upward only, p is at least as high as under the approval probability of any compliant state under $\tau\phi$. Next, we show noncompliant states pass with lower probability under $\omega_{p,\hat{s}}$. For illustration, let $\{t_n\}$ be a sequence of compliant states that converges to $\hat{s}$ and such that the sequence $p_n = \tau(t_n)$ converges to p. Such a sequence exists, because $\hat{s} \in \bar{S}^+$. Then, for every noncompliant state s, and every n, $\sup_t \ \tau(t) - \gamma c(t|s) \geq p_n - \gamma c(t_n|s)$. Going to the limit in n implies

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

which proves the point.

Since noncompliant states are approved with lower probability, and compliant states with higher probability, the receiver is better off under $\omega_{p,\hat{s}}$ than under ω. □

S2 Overt falsification: omitted results, proofs

S2.1 Equilibrium definition in the overt case

Note our definitions and notations in the modeling section carry on to any setup with a state space $S \subseteq \left[-\underline{s}, \overline{s}\right]$. Next, we add a few notations and provide a formal definition of an equilibrium in the overt case.

In the overt case, an approval strategy is a falsification-contingent plan, which we can denote as a family $\left\{\alpha_{\phi'}\right\}_{\phi'}$, where, for each possible falsification strategy ϕ', $\alpha_{\phi'}$ is a Markov kernel from X to A. The natural equilibrium concept is simply subgame-perfect equilibrium. Given a test τ, a strategy profile $\left(\phi, \left\{\alpha_{\phi'}\right\}_{\phi'}\right)$ is an equilibrium if:

(i) For every ϕ', $\alpha_{\phi'}(x) = \delta_a$ if $\mu(x|\tau\phi') > 0$, and $\alpha_{\phi'}(x) = \delta_r$ if $\mu(x|\tau\phi') < 0$,

(ii) For every ϕ', $U(\alpha_\phi \tau\phi, \phi) \geq U(\alpha_{\phi'}\tau\phi', \phi')$.

To ensure the agent's payoff is upper semicontinuous in ϕ, we can break the receiver's indifference in his favor, and assume she approves when her posterior mean is 0 that is, we can set the strategy of the receiver to $\{\bar{\alpha}_{\tau\phi'}\}_{\phi'}$, where

$$\bar{\alpha}_{\tau\phi'}(x) = \begin{cases} \delta_a & \text{if } \mu(x|\tau\phi') \geq 0 \\ \delta_r & \text{if } \mu(x|\tau\phi') < 0 \end{cases}.$$

Then, we say ϕ is overt equilibrium-feasible under τ, or that the pair (τ, ϕ) is overt equilibrium-feasible if

$$\forall \phi',\ U(\bar{\alpha}_{\tau\phi}\tau\phi, \phi) \geq U(\bar{\alpha}_{\phi'}\tau\phi', \phi'). \tag{OEF}$$

S2.2 Falsification-proofness principle

Although we only use the falsification-proofness principle with two states for overt falsification in the paper, it also holds for covert falsification. The following proposition presents its most general version. We also provide an example to illustrate why having more than two states may lead the principle to fail.

Proposition S2.1 (Falsification-proofness principle)**.** *Suppose falsification is costless or the state space is binary. Then, if (τ, ϕ) is (covert or overt) equilibrium feasible, the test $\tau' = \tau\phi$ and the truth-telling strategy δ also form an equilibrium* *feasible pair, with the same approval strategy α in the unobservable case. Furthermore, by definition, $\overline{V}(\tau\phi) = \overline{V}(\tau'\delta)$ and $\Pi(\alpha\tau'\delta) = \Pi(\alpha\tau\phi)$.*

Proof. **Overt case.** First consider a test τ and a falsification strategy ϕ such that (τ, ϕ) is overt equilibrium-feasible. Consider replacing τ by the test $\tau' = \tau\phi$. Then, the payoff of truth-telling under τ' is $U(\bar{\alpha}_{\tau\phi}\tau\phi, \delta) = \Pi(\bar{\alpha}_{\tau\phi}\tau\phi)$, and the payoff of using any other falsification strategy ϕ' is

$$\begin{aligned}\Pi(\bar{\alpha}_{\tau\phi\phi'}\tau\phi\phi') - C(\phi') &= U(\bar{\alpha}_{\tau\phi\phi'}\tau\phi\phi', \phi\phi') + C(\phi\phi') - C(\phi') \\ &\leq U(\bar{\alpha}_{\tau\phi}\tau\phi, \phi) + C(\phi\phi') - C(\phi') \\ &= \Pi(\bar{\alpha}_{\tau\phi}\tau\phi) - C(\phi) + C(\phi\phi') - C(\phi').\end{aligned}$$

Therefore, a sufficient condition for $(\tau\phi, \delta)$ to be equilibrium feasible is that $C(\phi\phi') \leq C(\phi) + C(\phi')$.

Covert case. Next, consider a test τ and suppose (ϕ, α) is an equilibrium with covert falsification under τ. Consider replacing τ by the test $\tau' = \tau\phi$ and keeping the same approval strategy α. Note that, under this new test, α is a best-response of the receiver to the agent telling the truth. Fixing α, the payoff of telling the truth under this new test is given by $U(\alpha\tau\phi, \delta) = \Pi(\alpha\tau\phi)$, and the payoff of deviating to ϕ' when δ is anticipated is given by

$$\begin{aligned}\Pi(\alpha\tau\phi\phi') - C(\phi') &= U(\alpha\tau\phi\phi', \phi\phi') + C(\phi\phi') - C(\phi') \\ &\leq U(\alpha\tau\phi, \phi) + C(\phi\phi') - C(\phi') \\ &= \Pi(\alpha\tau\phi) - C(\phi) + C(\phi\phi') - C(\phi'),\end{aligned}$$

where the inequality is due to the hypothesis that (ϕ, α) is an equilibrium under τ. Therefore, we obtain the same sufficient condition for $(\tau\phi, \delta)$ to be equilibrium feasible as in the observable case.

Satisfying the sufficient condition. The sufficient condition is trivially satisfied when falsification is costless. In the binary-state case, letting $\underline{c} = \gamma c(\overline{s}| - \underline{s})$, $\overline{c} = \gamma c(-\underline{s}|\overline{s})$, and, for any falsification strategy ϕ, $\underline{\phi} = \phi(\overline{s}| - \underline{s})$, and $\overline{\phi} = \phi(-\underline{s}|\overline{s})$, we have

$$\begin{aligned}C(\phi\phi') &= \pi\underline{c}\big\{\underline{\phi}'(1 - \overline{\phi}) + (1-\pi)\underline{\phi}(1 - \underline{\phi}')\big\} + \overline{c}\big\{\overline{\phi}(1 - \overline{\phi}') + \overline{\phi}'(1 - \underline{\phi})\big\} \\ &\leq \pi\underline{c}\big\{\underline{\phi}' + (1-\pi)\underline{\phi}\big\} + \overline{c}\big\{\overline{\phi} + \overline{\phi}'\big\} = C(\phi) + C(\phi').\end{aligned}$$

□

Remark 1. When it applies, our falsification-proofness principle resembles the standard revelation principle that applies when falsification is costless, but differs in subtle ways. Indeed, whereas the outcome $\tau\phi\delta$ is the same as $\tau\phi$, and the receiver's payoff is therefore the same under (τ, ϕ) and under $(\tau\phi, \delta)$, the agent's payoff is higher under $(\tau\phi, \delta)$ because he saves on falsification costs. A test designer can therefore restrict attention to falsification-proof tests whenever her objective only depends on the outcome, or is nondecreasing in the payoffs of the agent and the receiver. ◇

Example 1 (Failure of the falsification-proofness principle)**.** *The reason the sufficient condition $C(\phi\phi') \leq C(\phi) + C(\phi')$ generally fails with more than two states can be easily understood in a three-state example. Suppose $S = \{-3, 1, 3\}$, and ϕ falsifies 1 as 3 with certainty, whereas ϕ' falsifies -3 as 1 with certainty. Then, $\phi\phi'$ consists of falsifying both -3 and 1 as 3. Under any monotonic cost function, falsifying -3 as 3 is costlier than falsifying -3 as 1. Therefore, the sufficient condition is violated. Next, we illustrate how that can indeed lead to an optimal test that is not falsification-proof.*

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

S2.3 Normalization of signals as means

Consider a test τ. The corresponding cdf of conditional means in the absence of falsification is given by

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

To this cdf, we can associate a unique test $\hat{\tau}$ with signal space $\hat{X} = [-\underline{s}, \overline{s}]$ such that the cdf of conditional means generated by $\hat{\tau}$ is also H. This test is the one that pools together all signals that lead to the same posterior mean under τ, and relabels the pooled signal as this common mean. It is unique because $\hat{\tau}$ is characterized by $\hat{\tau}\big(\{x \in \hat{X} : x \leq y\}|\overline{s}\big) = \overline{H}(y)$, and $\hat{\tau}\big(\{x \in \hat{X} : x \leq y\}| -\underline{s}\big) = \underline{H}(y)$, where $\overline{H}$ and $\underline{H}$ are respectively given by ($\overline{\text{CDF}}$), and ($\underline{\text{CDF}}$). Then, $\hat{\tau}$ is the normalization by the mean of τ.

Lemma S2.1 (Normalization of signals as means). *For any falsification strategy ϕ', $\hat{\tau}$ leads to the same interim approval probabilities τ: $\overline{\alpha}_{\hat{\tau}\phi'}\hat{\tau}\phi'(s) = \overline{\alpha}_{\tau\phi'}\tau\phi'(s)$; and generates the same payoffs for the agent and the receiver: $U(\tau\phi', \phi') = U(\hat{\tau}\phi', \phi')$ and $V(\tau\phi') = V(\hat{\tau}\phi')$. In particular, if (τ, ϕ) is an equilibrium-feasible information structure, so is $(\hat{\tau}, \phi)$.*

Proof. We start by noting that, for every $\hat{x} \in \hat{X}$, $\hat{\tau}$ is just pooling together all signals $x \in X$ such that $\mu(x|\tau) = \hat{x}$, and relabelling the corresponding signal as $\hat{x}$. Let $X(\hat{x}) = \{x \in X : \mu(x|\tau) = \hat{x}\}$ be the set of signals that are pooled together. Then, for each $\hat{x} < \overline{s}$ and each $x \in X(\hat{x})$, the likelihood ratio $\lambda(x;\tau)$ informally defined as $\frac{\tau(dx|\overline{s})}{\tau(dx|-\underline{s})}$ exists and, by Bayes law, must satisfy

$$\lambda(x;\tau) = \frac{(1-\pi)(\hat{x}+\underline{s})}{\pi(\overline{s}-\hat{x})}.$$

Given that the right-hand side does not depend on x, the ratio is fully determined by the posterior mean $\hat{x}$ associated with signal x.

With falsification, we obtain:

$$\lambda(x;\tau\phi') = \frac{\tau(dx|\overline{s})}{\phi'\tau(dx|\overline{s}) + (1-\phi')\tau(dx|-\underline{s})(x)} = \frac{\lambda(x;\tau)}{\phi'\lambda(x;\tau)+1-\phi'}$$

for all $x \in X(\hat{x})$. Again, the likelihood ratio only depends on $\hat{x}$, and therefore, we also have $\mu(x|\tau\phi') = \mu(\hat{x}|\hat{\tau}\phi')$. In particular, a signal $x \in X(\hat{x})$ is approved under

$\tau\phi'$ if and only if $\hat{x}$ is approved under $\hat{\tau}\phi'$. Because, for any ϕ' and any $s \in S$, we have $\tau\phi'\big(X(\hat{x})|s\big) = \hat{\tau}\phi'(\hat{x}|s)$, the rest follows easily. □

S2.4 Other proofs

Proof of Lemma 2. Using the formulas from the proof of Lemma S2.1, the likelihood ratio informally defined by $\lambda(x) = \frac{d\overline{H}(x)}{d\underline{H}(x)}$ exists for every $x < \overline{s}$ and satisfies

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

which is strictly increasing in x. With falsification, this likelihood ratio is also well-defined and satisfies (again, from the proof of Lemma S2.1)

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

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

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

implying the threshold is $\overline{s}$, whenever $\frac{1}{\underline{\phi}} \leq \lambda(0)$, that is, whenever $\underline{\phi} \geq \varphi_0$. Otherwise, the threshold is equal to the unique x that solves $\lambda(x, \underline{\phi}) = \lambda(0)$. A bit of algebra then yields our formula for $\hat{x}(\underline{\phi})$, and the remaining claims are trivial. □

Proof of Proposition 4. The only part that needs additional explanations is the calculation of the agent's payoff. For illustration, note the receiver's payoff is

$$U\big(\mathcal{H}, \hat{\phi}(x)\big) = 1 - (\pi + (1-\pi)\hat{\phi}(x))\overline{H}_\ell(x) + (1-\pi)(1-\hat{\phi}(x))\underline{H}_\ell(x) - (1-\pi)\underline{c}\hat{\phi}(x).$$

The rest is algebra using formulas ($\overline{\text{CDF}}$) and ($\underline{\text{CDF}}$), as well as the identity $\mu_\pi = \pi\overline{s} - (1-\pi)\underline{s}$. □

Proof of Proposition 5. We have already proved $\mathcal{H}^*_{\underline{c}}$ is continuously differentiable and admits a density on $x > 0$, which is given by (2). Differentiating (2), we get

$$h^{*\prime}_{\underline{c}}(x) = \frac{h^*_{\underline{c}}(x)}{(x+\underline{s})(x-\mu_\pi)}(\mu_\pi - \underline{s} - x) < 0.$$

Differentiating the expressions in ($\overline{\text{CDF}}$) and ($\underline{\text{CDF}}$), we obtain that the densities of the belief distributions generated by the two types on $x > 0$ are

$$\overline{h^*_{\underline{c}}}(x) = \frac{x+\underline{s}}{(\mu_\pi+\underline{s})} h^*_{\underline{c}}(x),$$

and

$$\underline{h^*_{\underline{c}}}(x) = \frac{\overline{s}-x}{\overline{s}-\mu_\pi} h^*_{\underline{c}}(x).$$

A quick calculation yields

$$\overline{h^{*\prime}_{\underline{c}}}(x) = \frac{h^*_{\underline{c}}(x)}{(\mu_\pi+\underline{s})}\frac{-\underline{s}}{(x-\mu_\pi)} < 0,$$

and

$$\underline{h^{*\prime}_{\underline{c}}}(x) = \frac{h^*_{\underline{c}}(x)}{(\overline{s}-\mu_\pi)}\frac{\overline{s}\mu_\pi - \overline{s}\underline{s} - \overline{s}x + \mu_\pi\underline{s}}{(x-\mu_\pi)(x+\underline{s})} < 0.$$

To prove first-order stochastic dominance, we can use the expressions in ($\overline{\text{CDF}}$) and ($\underline{\text{CDF}}$) to get

$$\overline{H^*_{\underline{c}}}(x) - \underline{H^*_{\underline{c}}}(x) = \frac{\overline{s}+\underline{s}}{(\mu_\pi+\underline{s})(\overline{s}-\mu_\pi)}\Big\{(x-\mu_\pi)H^*_{\underline{c}}(x) - \mathcal{H}^*_{\underline{c}}(x)\Big\}.$$

Convexity of $\mathcal{H}$ implies this expression is negative for $x \geq 0$. For $x < 0$, we have $H^*_{\underline{c}}(x) = \kappa^*_{\underline{c}}$, and $\mathcal{H}^*_{\underline{c}}(x) = \kappa^*_{\underline{c}}x$ therefore,

$$\overline{H^*_{\underline{c}}}(x) - \underline{H^*_{\underline{c}}}(x) = -\frac{\overline{s}+\underline{s}}{(\mu_\pi+\underline{s})(\overline{s}-\mu_\pi)}\mu_\pi\kappa^*_{\underline{c}} < 0.$$

□

Proof of Proposition 6.

Comparative statics with respect to $\underline{c}$. Note (1) implies that for $x \in (\underline{s}, 0)$, $\mathcal{H}^*_{\underline{c}}(x) > \mathcal{H}^*_0(x)$, as it is easy to see that $\zeta(x) > 1$ (ζ and χ, below, are defined in the proof of Theorem 2 in the paper, and equation references also point to this proof). Using (1), and the functions we defined in step 1, we can write that, for

$x > 0$,

$$\mathcal{H}^*_{\underline{c}}(x) = \frac{\zeta(x)\big(1+\chi(x)\big)}{\zeta(\overline{s})\big(1+\chi(\overline{s})\big)}(\overline{s}-\mu_\pi)+\theta\underline{c}\big(1+\chi(x)\big)\zeta(x)\underbrace{\left(\frac{\big(\zeta(\overline{s})-1\big)}{\zeta(\overline{s})\big(1+\chi(\overline{s})\big)}-\frac{\zeta(x)-1}{\big(1+\chi(x)\big)\zeta(x)}\right)}_{A(x)}.$$

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

$$\begin{aligned}
\left(\frac{\zeta(x)-1}{\big(1+\chi(x)\big)\zeta(x)}\right)' &\propto (1+\chi(x))\zeta'(x)-\chi'(x)\zeta(x)\big(\zeta(x)-1\big)\\
&= \frac{x}{(x-\mu_\pi)(x+\underline{s})}\big(1+\chi(x)\big)\zeta(x)-\frac{1}{x+\underline{s}}\big(\zeta(x)-1\big)\\
&\propto \big(\mu_\pi+x\chi(x)\big)\zeta(x)+(x-\mu_\pi)\\
&= \left\{\frac{\mu_\pi}{\underline{s}}(x+\underline{s})+x\left(\frac{-\mu_\pi}{\underline{s}}\right)^{\frac{\mu_\pi}{\mu_\pi+\underline{s}}}\left(\frac{x-\mu_\pi}{x+\underline{s}}\right)^{\frac{\underline{s}}{\mu_\pi+\underline{s}}}\right\}\left(\frac{x-\mu_\pi}{-\mu_\pi}\right)^{\frac{\mu_\pi}{\mu_\pi+\underline{s}}}\\
&\qquad\times\left(\frac{x+\underline{s}}{\underline{s}}\right)^{\frac{\underline{s}}{\mu_\pi+\underline{s}}}+(x-\mu_\pi)\\
&= \mu_\pi\left(\frac{x-\mu_\pi}{-\mu_\pi}\right)^{\frac{\mu_\pi}{\mu_\pi+\underline{s}}}\left(\frac{x+\underline{s}}{\underline{s}}\right)^{1+\frac{\underline{s}}{\mu_\pi+\underline{s}}}+\frac{x+\underline{s}}{\underline{s}}(x-\mu_\pi)\\
&\propto -\left(\frac{x-\mu_\pi}{-\mu_\pi}\right)^{-\frac{\underline{s}}{\mu_\pi+\underline{s}}}\left(\frac{x+\underline{s}}{\underline{s}}\right)^{1+\frac{\underline{s}}{\mu_\pi+\underline{s}}}+\frac{x+\underline{s}}{\underline{s}}\\
&= \left(\frac{x+\underline{s}}{\underline{s}}\right)\left\{1-\left(\frac{-\mu_\pi(x+\underline{s})}{(x-\mu_\pi)\underline{s}}\right)^{\frac{\underline{s}}{\mu_\pi+\underline{s}}}\right\}\\
&> 0,
\end{aligned}$$

where the last inequality is obtained by noticing the first term is positive and the second term is decreasing in x and therefore bounded below by its value at $x = \overline{s}$, which is positive because $\frac{\overline{s}+\underline{s}}{\underline{s}} > \frac{\overline{s}-\mu_\pi}{-\mu_\pi} \Leftrightarrow \mu_\pi > -\underline{s}$.

This finding shows that for every $x \in (0, \overline{s})$, $\mathcal{H}^*_{\underline{c}}(x)$ is increasing in $\underline{c}$ and furthermore $\mathcal{H}^*_{\underline{c}}(x) > \mathcal{H}^*_0(x)$. The same holds on $(-\underline{s}, 0]$ by (1). This argument proves the comparative statics with respect to the Blackwell informativeness ordering. The comparative statics for the receiver's payoff also follows.

$\mathcal{H}^*_{\underline{c}}$ is more informative than any other receiver-optimal test. First, if $\mathcal{H}$ is another receiver-optimal test, we can linearize it to the left of 0, which makes it more informative. Next, suppose that, for some $\hat{x} \in (0, \overline{s})$, $\mathcal{H}(\hat{x}) > \mathcal{H}^*_{\underline{c}}(\hat{x})$. Then, we can replicate the optimality argument of step 3 in the proof of Theorem 2 to find a contradiction. Therefore, for all $x \in (0, \overline{s})$, we have $\mathcal{H}(x) \leq \mathcal{H}^*_{\underline{c}}(x)$. Because the two test functions must be equal to the left of 0 because they are linear and deliver the same receiver payoff, we can conclude $\mathcal{H}$ is less informative than $\mathcal{H}^*_{\underline{c}}$.

□

Proof of Proposition 7.

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

Payoff bound. The full-information payoff of the receiver is $\pi\overline{s} = \frac{\underline{s}+\mu_\pi}{\overline{s}+\underline{s}}\overline{s}$. First, to obtain a lower bound on the payoff ratio, note the three-signal test we obtained in Section 4.1 yields a payoff equal to $\pi\overline{s}\left(\frac{\underline{s}+\overline{s}}{\underline{s}+2\overline{s}}\right) \geq \frac{1}{2}\pi\overline{s}$ in the absence of cost. Because our optimal test does better, it delivers more than one half of the full-information payoff in the absence of falsification costs, and yet more with a positive cost. Next, we show the bound is tight in the absence of cost. Note the payoff ratio can be written as

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

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

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

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

□

Proof of Proposition 8. First, $\mathcal{H}^*_{\underline{c}}$ is immune against any deviation such that $\overline{\phi}+$

$\underline{\phi} \leq 1$. To prove that we show that deviating to a falsification strategy $(\underline{\phi}, \overline{\phi})$ such that $\underline{\phi} + \overline{\phi} \leq 1$ is dominated by the strategy $(\underline{\phi}, 0)$. It leads the receiver to use a threshold $\hat{x} \geq \hat{x}\big(\underline{\phi}\big)$ because the probability that a signal in the continuum is generated by the compliant state is lower than under $(\underline{\phi}, 0)$, whereas the probability that it is generated by the noncompliant state is the same. Furthermore, lowering the probability that the compliant state generates a passing signal is also harmful in itself. Because $(\underline{\phi}, 0)$ is, by construction, not profitable, the same is true for $(\underline{\phi}, \overline{\phi})$.

Next, any deviation such that $\overline{\phi} + \underline{\phi} > 1$ devalues the mean associated with all signals in the continuum of passing signals below 0, so they all lead to rejection, whereas the mean associated with the unique signal rejected in the original test is proportional to $\overline{\phi}\pi\overline{s} - (1 - \underline{\phi})(1 - \pi)\underline{s}$. The mean is therefore nonnegative if and only if $1 - \underline{\phi} \leq \varphi_0\overline{\phi}$, making any falsification strategy that does not satisfy this inequality dominated. For falsification strategies that satisfy this inequality, the agent's payoff is given by

$$\pi\overline{\phi}(1 - \overline{c}) + (1 - \pi)\big(1 - \underline{\phi}(1 + \underline{c})\big).$$

This payoff is decreasing in $\underline{\phi}$, so the agent will always choose $1 - \underline{\phi} = \varphi_0\overline{\phi}$, yielding a payoff of $\overline{\phi}\big(\pi(1 - \overline{c}) + (1 - \pi)\varphi_0(1 + \underline{c})\big) - \underline{c}(1 - \pi)$. This payoff can be positive only if $\overline{\phi} = 1$, and because the agent can always grant himself a positive payoff by not falsifying, we need only consider the deviation $\overline{\phi} = 1$ and $\underline{\phi} = 1 - \varphi_0$.

Then, the test $\mathcal{H}^*_{\underline{c}}$ remains optimal if and only if this deviation is not profitable to the agent, that is, if and only if

$$1 - \kappa^*_{\underline{c}} \geq \pi(1 - \overline{c}) + (1 - \pi)\big(\varphi_0 - (1 - \varphi_0)\underline{c}\big) = \pi\left(\frac{\overline{s} + \underline{s}}{\underline{s}} - \overline{c}\right) + \frac{\mu_\pi}{\underline{s}}\underline{c}.$$

Replacing $\kappa^*_{\underline{c}}$ by its expression, and letting $\Lambda = \left(\frac{\overline{s} - \mu_\pi}{-\mu_\pi}\right)^{\frac{\mu_\pi}{\mu_\pi + \underline{s}}} \left(\frac{\overline{s} + \underline{s}}{\underline{s}}\right)^{\frac{\underline{s}}{\mu_\pi + \underline{s}}}$, we get the necessary and sufficient condition

$$(\underline{s} + \mu_\pi)\Lambda - \theta\underline{c}(\Lambda - 1) \geq \left(\pi\left(\frac{\overline{s} + \underline{s}}{\underline{s}} - \overline{c}\right) + \frac{\mu_\pi}{\underline{s}}\underline{c}\right)\big(\overline{s} - \mu_\pi + (\underline{s} + \mu_\pi)\Lambda\big).$$

A bit of algebra then yields the condition $A\overline{c} + B\underline{c} \geq 1$, where

$$A = \frac{\underline{s}(1 - \pi + \pi\Lambda)}{(1 - \pi)(\overline{s} + \underline{s}) + \Lambda\mu_\pi} > 0,$$

and

$$B = \frac{\frac{1-\pi}{\pi}(\underline{s} - \pi\overline{s}) - \Lambda\pi\overline{s}}{(1-\pi)(\overline{s}+\underline{s}) + \mu_\pi\Lambda}.$$

The positivity of A is implied by the fact that

$$\frac{-\Lambda\mu_\pi}{(1-\pi)(\overline{s}+\underline{s})} = \Lambda\frac{-\mu_\pi}{\overline{s}-\mu_\pi} = \left(\frac{(\overline{s}+\underline{s})(-\mu_\pi)}{(\overline{s}-\mu_\pi)\underline{s}}\right)^{\frac{\underline{s}}{\mu_\pi+\underline{s}}} < 1,$$

as $\frac{\underline{s}}{\mu_\pi+\underline{s}} > 1$, and $\frac{(\overline{s}+\underline{s})(-\mu_\pi)}{(\overline{s}-\mu_\pi)\underline{s}} < 1$. □

S3 Complements on the binary-state model

S3.1 Optimal three-signal test

We show the three-signal test described in the paper is in fact optimal among three-signal tests.

Proposition S3.1. *The optimal three-signal test is*

$$\mathcal{H}^*_{3S}(x) = \kappa^*_{3S}(x+\underline{s}) + \frac{\overline{s} - \mu_\pi - \kappa^*_{3S}(\underline{s}+\overline{s})}{\overline{s}}\max\{x, 0\},$$

which coincides with the three-signal test described in the paper.

Proof. First, one of the three signals must be at 0 for the test to be falsification-proof. Second, the same linearization argument as for the optimal test shows another signal must be at $-\underline{s}$. Therefore, the only unknown is the position of the third signal $x \in (0, \overline{s}]$. Using the linearization, we denote by κ the slope of the test function $\mathcal{H}$ between signals $-\underline{s}$ and 0. The slope η of the test function between 0 and x must satisfy

$$\mathcal{H}(x) = \kappa\underline{s} + \eta x = x - \mu_\pi$$

hence,

$$\eta = \frac{x - \mu_\pi - \kappa\underline{s}}{x}.$$

The no-falsification incentive constraint of the agent, adapted from (FPIC'), then states

$$\eta - \frac{x}{\underline{s}+x} \geq \frac{\underline{s}}{\underline{s}+x}\kappa - \frac{\theta\underline{c}x}{(x-\mu_\pi)(\underline{s}+x)}.$$

Replacing η by its value, we obtain the constraint

$$\kappa \leq \frac{1}{\underline{s}} \left\{ \frac{x(\underline{s}-\mu_\pi) - \mu_\pi \underline{s}}{2x+\underline{s}} + \theta \underline{c} \frac{x^2}{(2x+\underline{s})(x-\mu_\pi)} \right\}.$$

The program of the designer is to maximize $\mathcal{H}(0) = \kappa \underline{s}$, hence κ, under this constraint. It is easy to verify that the right-hand side of the constraint is increasing in x. Therefore, setting $x = \overline{s}$ is optimal, and

$$\kappa^*_{3S} = \frac{1}{\underline{s}} \left\{ \frac{\overline{s}(\underline{s}-\mu_\pi) - \mu_\pi \underline{s}}{2\overline{s}+\underline{s}} + \theta \underline{c} \frac{\overline{s}^2}{(2\overline{s}+\underline{s})(\overline{s}-\mu_\pi)} \right\}.$$

Recalling that $\theta = \frac{(\overline{s}-\mu_\pi)(\underline{s}+\mu_\pi)}{(\underline{s}+\overline{s})}$, that is

$$\kappa^*_{3S} = \frac{1}{\underline{s}} \left\{ \frac{\overline{s}(\underline{s}-\mu_\pi) - \mu_\pi \underline{s}}{2\overline{s}+\underline{s}} + \underline{c} \frac{(\underline{s}+\mu_\pi)\overline{s}^2}{(2\overline{s}+\underline{s})(\overline{s}+\underline{s})} \right\}$$

Note κ^*_{3S} is the ex ante probability of generating the lowest signal, which is $(1-\pi)(1-\underline{p})$ under the three-signal test from the paper. A straightforward calculation shows these probabilities are equal, so the two tests are identical. □

S3.2 Overt-covert comparison

We start by deriving optimal tests under covert falsification. In the binary-state case, with covert falsification, we can use both the falsification-proofness principle, by Proposition S2.1, and the recommendation principle, by Lemma S1.1. Using the recommendation principle, we denote the test by $\tau = (\underline{\tau}, \overline{\tau})$, where $\underline{\tau}$ is the nominal passing probability of the low state $-\underline{s}$, and $\overline{\tau}$ that of the high state. Then, the set of equilibrium-feasible approval probabilities is characterized by the obedience constraint

$$\overline{\tau}\pi\overline{s} - \underline{\tau}(1-\pi)\underline{s} \geq \max\{\mu_\pi, 0\} \tag{S3.1}$$

and the falsification proofness constraint[1]

$$\overline{\tau} - \underline{\tau} \leq \underline{c}, \tag{S3.2}$$

[1] It is easy to show (S3.1) implies that the second falsification-proofness constraint $\underline{\tau} - \overline{\tau} \leq \overline{c}$ is redundant.

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

Suppose $\mu_\pi < 0$. Then, the uninformative and obedient test $\tau_{NI} = (0,0)$ is pessimal for both players; the fully informative and obedient test $\tau_{FI} = (0,1)$ yields the first best for the receiver, whereas the agent optimal obedient test is $\tau_{KG} = (\varphi_0, 1)$, where KG stands for Kamenica-Gentzkow because this information structure is agent (aka sender) optimal. When $\underline{c} \geq 1$, all these tests also satisfy (S3.2), and the set of equilibrium-feasible information structures is $\operatorname{co}(\{\tau_{NI}, \tau_{FI}, \tau_{KG}\})$, which coincides with what is feasible without falsification. At the other extreme, when $\underline{c} = 0$, only $\tau_{NI} = (0,0)$ is feasible. We now turn to the interesting range of falsification costs $\underline{c} \in (0,1)$. Elementary algebra yields that $\tau_R = (0, \underline{c})$ is the receiver-optimal test. Coming to the agent, the range $\underline{c} \in (0,1)$ can be divided into two regions depending on whether τ_{KG} is feasible. By construction, τ_{KG} satisfies (S3.1) with equality, but it violates (S3.2) when $\underline{c} \leq 1 - \varphi_0$; in this range, the agent-optimal test is the one that satisfies both (S3.1) and (S3.2) with equality, $\tau_A = \left(-\frac{\underline{c}\pi\overline{s}}{\mu_\pi}, \frac{-\underline{c}(1-\pi)\underline{s}}{\mu_\pi}\right)$. When τ_{KG} satisfies (S3.2) with slack, which happens when $\underline{c} > 1 - \varphi_0$, another extremal information structure arises: the test $\tau_P = (1 - \underline{c}, 1)$ that satisfies (S3.2) with equality but (S3.1) with slack.[2] Then, the set of equilibrium-feasible tests is:

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

We depict $\mathcal{T}$ in Figure 1 for various cost levels. The corresponding set of feasible payoffs is depicted in Figure 2, where we compare it with the set of feasible payoffs under observable falsification.

With two states, we can rely on falsification-proof tests, so inefficiency occurs due to incurred costs. However, when $\underline{c} < 1$, the receiver-optimal test is informationally inefficient due to inefficient approval of the high state. Furthermore,

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

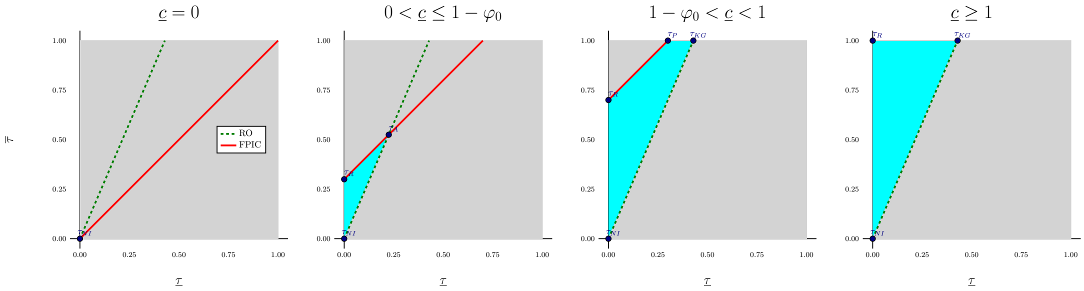

Figure 1: *Covert falsification: the blue region is the set of feasible information structures* $\mathcal{T}$. *Parameters for the plots:* $-\underline{s} = -2, \overline{s} = 2, \pi = 0.3, (\mu_\pi = -0.8)$; $\underline{c} \in \{0, 0.3, 0.7, 1\}$.

if $\underline{c} < 1 - \varphi_0$, no efficient (or informationally efficient) feasible test exists. If $\underline{c} \geq 1 - \varphi_0$, all tests on $\operatorname{co}\big(\{\tau_{KG}, \tau_P\}\big)$ are efficient. As we show next, in the continuous-state case, the receiver-optimal test is always inefficient due both to falsification costs incurred by the agent and to informational inefficiency for sufficiently low costs.

We finish by comparing the equilibrium outcomes arising under unobservable and observable falsification. In Figure 2, we depict feasible payoffs under overt and covert falsification, in the binary-state case. The set of feasible payoffs under covert falsification (in blue) is also achievable under overt falsification, because it is easy to see the agent has no incentive to falsify any of the tests at its extreme point under overt falsification. The test τ_{KG}, whose payoffs lie at the top vertex of the grey payoff triangle, is falsification-proof under overt falsification, and therefore feasible. Finally, our receiver-optimal test is also feasible. This finding implies all payoffs in the pink area are feasible under overt falsification. Furthermore, we know no payoff vector to the right of the receiver-optimal test payoff vector is feasible. Overall, these observations show that making falsification observable, or equivalently giving the agent the means to commit to his falsification strategy, enlarges the set of feasible payoffs, and makes attaining efficiency even when upward falsification is costless possible.

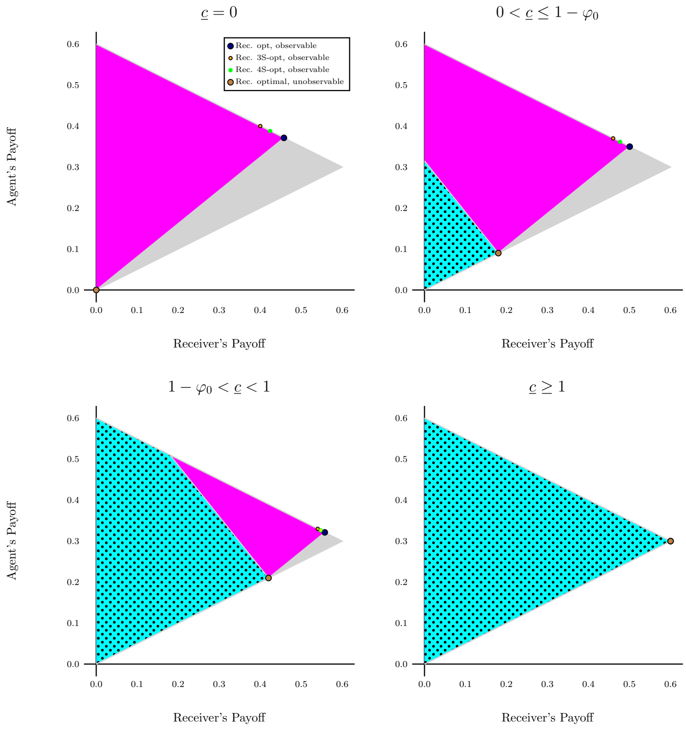

Figure 2: *The grey triangle depicts the space of feasible payoffs without falsification. The blue dotted area depicts the set of feasible payoffs under covert falsification. The pink area shows some of the additional payoffs that are feasible under overt falsification.*