Evidence Acquisition and Voluntary Disclosure*

Denis Shishkin †

June 4, 2024

(click here for the latest version).

Abstract

A sender seeks hard evidence to persuade a receiver to accept a project by designing a quality test. Testing is not perfectly reliable and produces evidence only with some probability. If the sender obtains the evidence, she can choose to disclose it or pretend to not have obtained it. We show that when reliability is low, the sender chooses a pass/fail test that reveals whether the quality is above or below a threshold. Moreover, the equilibrium pass/fail threshold is always monotone in reliability but whether it is increasing or decreasing depends on whether evidence acquisition is overt or covert.

*I am grateful to Stephen Morris and Pietro Ortoleva for their guidance in developing this project. I would like to thank Nageeb Ali, Roland Bénabou, Simone Galperti, Faruk Gul, Nima Haghpanah, Elliot Lipnowski, Konrad Mierendorff, Oleg Muratov, Franz Ostrizek, Wolfgang Pesendorfer, Doron Ravid, Evgenii Safonov, Vasiliki Skreta, Joel Sobel, Can Urgun, Nikhil Vellodi, Joel Watson, Leeat Yariv, and various seminar and conference audience members for helpful comments and insightful discussions.

†Department of Economics, University of California San Diego, dshishkin@ucsd.edu.

1 Introduction

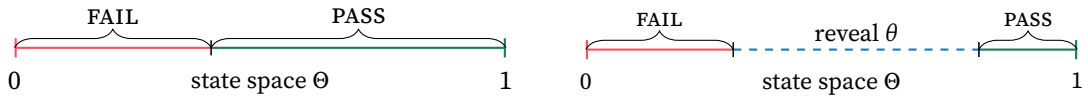
Hard evidence is often sought and disclosed by one party (sender) to persuade another (receiver) to take a certain action. For example, pharmaceutical companies test new drugs and seek approval from the US Food and Drug Administration, startups build and test prototypes to secure funding from investors, sellers certify quality of their products to persuade consumers to buy them, etc. However, in many cases the receiver may be uncertain about whether the sender has obtained the evidence. In the above examples, it could be that by the time of the final decision the testing results may not have come back or may have come back inconclusive. In this case, even if the sender has evidence, she might be able to pretend to not have obtained any evidence. In other words, she can conceal sufficiently unfavorable evidence by claiming ignorance. This paper studies the trade-off arising from the conflict between the sender's preferences over disclosures before and after she obtains the evidence.

In principle, when the state and message spaces are rich and information acquisition is costless, one might expect to see complex communication between the agents. In reality, however, senders often rely on coarse verifiable information. In many cases, it is as simple as a *pass/fail test*, that is, a signal that reveals only whether the state of the world is sufficiently good. This paper shows that the mere opportunity to conceal information as described above can lead to the design of very simple tests such as a pass/fail test in equilibrium.

To study these interactions, we consider a communication game between a sender (she) and a receiver (he). The state of the world is continuous and unknown to both players. The sender wants the receiver to take one of two actions, but the receiver does so only if his expectation of the state exceeds his privately known outside option drawn from a unimodal distribution. The sender chooses what hard information to acquire by designing an informative test about the state. However, such testing is not perfectly reliable, in particular, she obtains hard evidence about the test results only with some commonly known probability ρ, referred to as *reliability*. Even if she obtains the evidence, she can then voluntarily disclose it or pretend to not have obtained it. Otherwise, she cannot prove that she is uninformed. We distinguish between two versions of the model. In the case of *overt* acquisition, the sender's choice of information is always observed. In the case of *covert* acquisition, the

sender's choice of information is unobserved by the receiver unless evidence is disclosed. The covert case captures situations in which the sender cannot commit to the design of the test.

Our results (Theorems 1 and 2) characterize the equilibrium evidence structures in the overt and covert cases and show that they are essentially unique. The first key implication of the characterization is that low reliability leads to the simplicity of equilibrium evidence structures chosen by the sender. In particular, we show that if ρ is below a certain cutoff, the equilibrium structure takes the form of a *pass/fail test*: it reveals only whether the state is above or below some threshold. When ρ is instead above the cutoff, it takes the form of a two-sided censorship, which is similar to a pass/fail test, but also perfectly reveals some intermediate states (see Figure 1).



(a) Pass/fail test

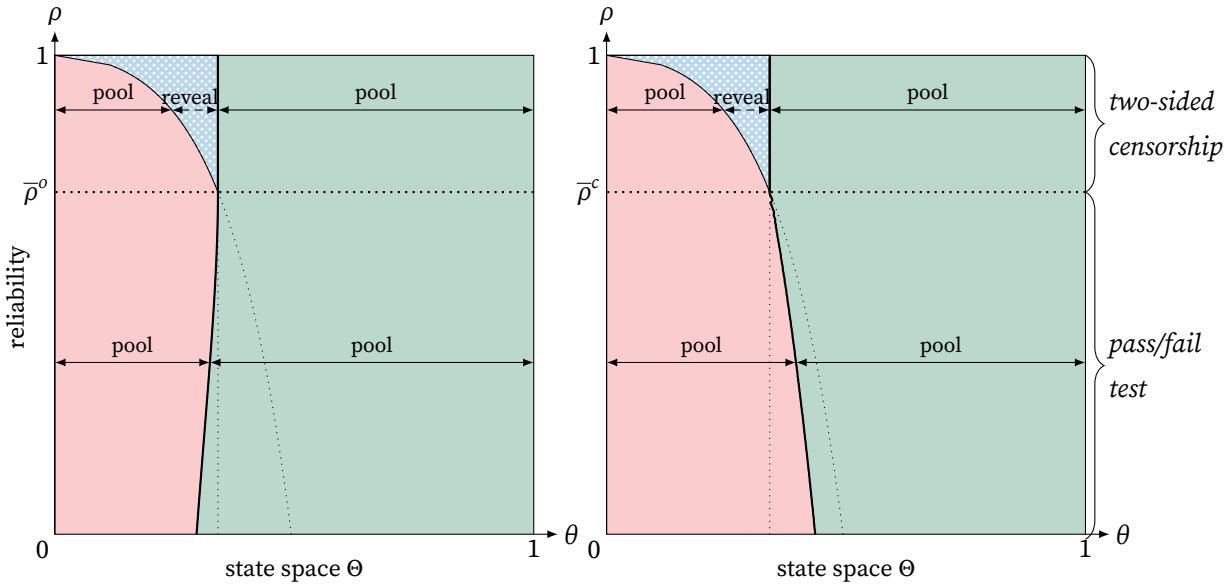
(b) Two-sided censorship

Figure 1: Two types of equilibrium evidence structures.

Second, we show that the equilibrium pass/fail threshold is monotone in reliability. However, the type of monotonicity depends on whether the nature of evidence acquisition: the pass/fail threshold is increasing in the overt case, but decreasing in the covert case. In other words, whether the sender publicly or privately acquires the evidence affects how testing standards react to improvements in reliability. In addition, we show that covert equilibrium pass/fail threshold is always strictly higher than the overt one and that the difference between them shrinks as reliability improves.

Figure 2 illustrates the key features of the equilibrium evidence structures for the uniformly distributed state and the receiver's outside option following the triangular distribution the peak at $3/5$. For each reliability level $\rho \in (0, 1]$, the corresponding horizontal line segment illustrates the optimal partition of the state space. When reliability is low ($\rho < \overline{\rho}$), there is a pass/fail threshold, such that states are pooled above and below it. When reliability is high ($\rho > \overline{\rho}$), the states are pooled above the upper threshold, pooled below the lower threshold, and fully revealed otherwise.

To get some intuition for these results, note that information is affected by three key



(a) Overt acquisition

(b) Covert acquisition

Figure 2: Equilibrium evidence structures for the uniformly distributed state and triangular distribution of the receiver's outside option with the peak at $3/5$.

forces arising due to voluntary disclosure, information design, and the (c)overt nature of the acquisition. First, because the sender does not want to reveal bad news, this prevents the receiver from learning detailed information about low states. Hence, the voluntary disclosure force drives the lower pooling region. Second, because the sender is uncertain about the receiver's cutoff for action and the distribution of the receiver outside options is unimodal, there are increasing returns to disclosing more (less) information about low (high) states. In particular, it is well known from the Bayesian persuasion literature[1] that a sender with full commitment and convex-concave indirect utility over posterior means will choose an *upper censorship* of the state, that is, a signal which reveals all states below a certain threshold and pools all states above it. In other words, the information design force drives the imprecision of information about high states which leads to the upper pooling the state.

Moreover, *whether* and *how* these two forces interact depends on the level of reliability and the third force—whether acquisition is covert. Note that for high reliability, only the lower pooling region is affected by reliability. In fact, in this case we show that the lower

[1] See, for example, Alonso and Câmara (2016a), Kolotilin (2018), Dworczak and Martini (2019).

threshold corresponds to the disclosure threshold obtained in the equilibrium of the voluntary disclosure game where the sender is fully informed (subject to reliability) and the upper threshold is constant and corresponds to the full commitment case (which coincides with our perfect reliability case). That is, for $\rho > \overline{\rho}$ the two thresholds are determined independently by the two forces and do not interact. Notably, the equilibria of the overt and covert cases coincide. Since, compared to the covert case, overt case essentially adds commitment to evidence structure (but not disclosure), this means that the corresponding additional incentive constraints of the covert case are slack. In other words, the overt equilibrium signal can resist potential ex-ante deviations by the sender which would not be detected under non-disclosure. We show that this is because the only additional benefit from an ex-ante covert deviation compared to the overt case might come from a lower non-disclosure receiver's posterior which is minimized under the equilibrium two-sided censorship.[2]

In contrast, under $\rho < \overline{\rho}$, the equilibrium signal is a pass/fail test and the threshold is determined jointly by the interaction between voluntary disclosure and information design. In the overt case, the receiver fully observes the sender's ex-ante choice and, hence, solving for equilibria boils down to an optimization problem. We show that such sender's costless acquisition problem that takes into account voluntary disclosure can be reformulated as a costly information design problem. That is, her ex-ante expected value from seeking an evidence structure is proportional to her perfect-reliability commitment value from actually choosing a distribution of R's posteriors minus the 'concealment loss' arising due to strategic non-disclosure of bad news. We first show that the solution to this costly information design problem shares similarities with the full-commitment ($\rho = 1$) case in that the solution must be disclosure-equivalent to an upper-censorship. Then, focusing on equilibria in which the sender does not acquire more information than needed given her strategic concealment yields a two-sided censorship or a pass/fail test depending on whether the disclosure threshold is above or below the upper pooling threshold. We then show that the concealment loss features substitutability between reliability and the testing standards implying that the overt equilibrium pass/fail threshold is increasing in

[2] As explained in Section 3.5, this is related to the minimum principle of DeMarzo, Kremer, and Skrzypacz (2019) which is both necessary and sufficient for covert equilibria in the case of uniformly distributed outside option.

reliability.

In the covert case, the sender's ex-ante choice is unobservable and, thus, solving for equilibria instead involves a fixed-point problem with respect to the sender's ex-ante choice of experiment π and the receiver's non-disclosure posterior $x_\varnothing$: (i) π must be best-responding to $x_\varnothing$ and (ii) $x_\varnothing$ must be Bayes-consistent given π. We show that the problem of finding the sender's best response to any given $x_\varnothing$ is equivalent to an auxiliary information design problem with the sender's indirect utility clipped at the bottom. Such a modification nevertheless leads again to an upper censorship solution. Finally, we show that the best-responding upper pooling threshold is increasing in $x_\varnothing$ and does not depend on reliability. At the same time, higher reliability leads to larger skepticism and hence to a lower Bayes-consistent non-disclosure posterior. This implies that the covert equilibrium pass/-fail threshold is decreasing in reliability.

We also study welfare implications of the observability of the sender's acquisition strategy. As discussed above, under high reliability ($\rho > \overline{\rho}$), the equilibria under overt and covert acquisition coincide and so both players are indifferent between the two cases. However, when reliability is low ($\rho < \overline{\rho}$), the equilibria are different between the two cases and the players are no longer indifferent. The seller is better off in the overt case due Stackelberg-like first-mover advantage: in the overt case, she has fewer incentive constraints. In other words, she is better off (in fact, strictly so, as we show) in the overt case because she could always replicate the covert equilibrium outcome. On the other hand, we show that the receiver strictly prefers the covert case. Despite the pass/fail tests with different thresholds being Blackwell-incomparable, we show that the information structure corresponding to the equilibrium sender-disclosed pass/fail tests is Blackwell-increasing in the threshold in the relevant range. And because the covert equilibrium pass/fail threshold is always strictly higher than the overt one, the receiver strictly prefers the former to the latter.

Related literature. This paper is related to the literature on disclosure of verifiable information (for a survey, see Milgrom, 2008). The seminal works of Grossman (1981), Milgrom (1981), and Milgrom and Roberts (1986) study disclosure under complete provability, meaning the sender can prove any true claim. The key insight of those papers is that complete provability implies "unraveling", which leads to full information revelation in equi-

librium.[3] Our model is based on the model of Dye (1985) also analyzed by Jung and Kwon (1988), in which evidence is obtained with some probability and there is partial provability: if the sender is uninformed, she cannot prove this.

The paper contributes to the literature that endogenizes the sender's endowment of evidence in voluntary disclosure games. In Matthews and Postlewaite (1985), the sender makes a binary evidence acquisition decision before playing a voluntary disclosure game under complete provability. Lizzeri (1999), Ali, Haghpanah, Lin, and Siegel (2021), and Asseyer and Weksler (2024) study disclosure of verifiable information designed by a profit-maximizing monopolistic intermediary. Gentzkow and Kamenica (2017) study overt, costly evidence acquisition in a disclosure model where each type can perfectly self-certify and show that one or more sender(s) disclose everything they acquire. Escudé (2024) provides an analogous result in a single-sender setting with covert costless acquisition and more general verifiability structures. Ben-Porath, Dekel, and Lipman (2021) study a mechanism design problem with privately informed agents who can acquire evidence about their types.

Some recent papers endogenize sender's evidence withing the Dye (1985) framework. Kartik, Lee, and Suen (2017) study a multi-sender disclosure game, where senders can invest in higher reliability, while taking the evidence structure as given. Dasgupta, Krasikov, and Lamba (2022) study hard information design in a monopolistic screening model. In Bertomeu, Cheynel, and Cianciaruso (2021), a regulator designs a firm's evidence structure subject to reliability and a cost of non-disclosure. Whitmeyer and Zhang (2022) study both overt and covert costly acquisition of evidence with an additional cost of disclosure.

The most closely related paper is DeMarzo et al. (2019) which studies a problem that can be related to our covert case with the uniform outside option and where the sender's choice can be across any constrained collection of tests of possibly heterogeneous reliability.[4] They show that a test arises in some equilibrium if and only it satisfies the 'minimum principle', that is, it must minimize the Bayes-consistent receiver's non-disclosure posterior. Notably, their results imply that there is always an equilibrium with 'simple tests' equivalent to our pass/fail tests. In contrast to our model, the information design force

[3] For a recent generalization, see Hagenbach, Koessler, and Perez-Richet (2014).

[4] Ben-Porath, Dekel, and Lipman (2018) study a related voluntary disclosure problem, in which there is an ex-ante covert choice between risky projects, which, in our setting, would translate into a choice between priors.

is absent in theirs because the sender's indirect utility over posterior means is linear and, therefore, she is ex-ante indifferent between all information structures.

This paper also contributes to the literature on Bayesian persuasion and information design (for a survey, see Kamenica, 2019). In the special case of our model when the sender is known to possess the evidence ($\rho = 1$), the unraveling argument applies, and both the overt and covert optimal evidence acquisition problems become equivalent to Bayesian persuasion (Aumann and Maschler, 1995; Kamenica and Gentzkow, 2011). In particular, a number of papers (Alonso and Câmara, 2016b; Kolotilin, Mylovanov, Zapechelnyuk, and Li, 2017; Kolotilin, 2018; Dworczak and Martini, 2019) have shown in similar settings that upper censorship is optimal if the receiver's type distribution is unimodal.[5] Information structures equivalent to our pass/fail test and two-sided censorship also appear in Kolotilin (2018) in cases when the distribution of the receiver's type is not unimodal. There, pass/fail test can be optimal because of a particular shape of the receiver's type distribution (e.g. bimodal), rather than the interaction between the design and disclosure incentives.

A standard assumption in this literature is that the sender commits to a signal, whose realization is directly observed by the receiver, while in our model it is voluntarily disclosed by the sender. [6] Some recent works also relax the assumption that the receiver directly observes signal realizations. In Felgenhauer (2019), the sender designs experiments sequentially at a cost and can choose when to stop experimenting and which outcomes to disclose. Nguyen and Tan (2021) study a model of Bayesian persuasion with costly messages, where a special case of the cost function corresponds to verifiable disclosure of hard evidence studied in this paper. The difference is that their sender can choose not only a signal about the state, but also the reliability. In contrast, ρ is exogenous in our model. If it could be chosen by the sender, she would set $\rho = 1$ and obtain her full-commitment payoff.

2 Model

Setup. There are two players: a sender (S, she) and a receiver (R, he). The state of the world is $\theta \in \Theta = [0, 1]$, unknown by both players, who share a prior belief with a CDF $\overline{F}$, a

[5] Moreover, Kolotilin, Mylovanov, and Zapechelnyuk (2022) establish the converse.

[6] See also Onuchic (2024) for a model in the sender can commit to a disclosure rule for realizations of an exogenously given signal.

full-support density $\bar{f}$ and a mean x_0. R has a privately known outside option $\omega \in \Omega = [0,1]$ drawn from some unimodal distribution independent of θ. In particular, assume that its CDF G admits a strictly quasiconcave full-support density g with a peak at some $\hat{\omega} > x_0$.[7] R has two actions: accept ($a = 1$) and reject ($a = 0$). The players's utility functions are given by $u_R(a, \theta, \omega) = a(\theta - \omega)$ and $u_S(a) = a$. That is, R prefers to accept if and only if his expectation of the state is at least as high as his outside option and S always wants R to accept.

The timing of the game is as follows.

1. S decides which evidence to seek. Formally, S chooses a *test*, i.e., a measurable mapping $\pi \colon \Theta \to \Delta M$, where $M = [0,1]$ is the message space.[8]
2. Nature draws an outside option ω from G, a state θ from $\overline{F}$, a message m from $\pi(\theta)$, and the set of available messages $\hat{M}$ as follows:
   - With probability $\rho \in (0,1]$, $\hat{M} = \{m, \varnothing\}$ which is interpreted as S obtaining a proof that the realized message is m;
   - With probability $1 - \rho$, $\hat{M} = \{\varnothing\}$, which can be interpreted either as S not being able to prove which outcome realized or that S has not learned the outcome of the experiment at all.
3. S observes $\hat{M}$ and chooses $\hat{m} \in \hat{M}$. That is, even if S obtains evidence, she can choose whether to disclose it or claim to not have obtained it.[9]
4. We distinguish between two variants of the game, depending on whether the evidence structure chosen by S is observed by R:
   - Under *covert evidence acquisition*, R observes $\hat{m}$ and, if $\hat{m} \neq \varnothing$, also observes π. Then he updates his belief and chooses an action;
   - Under *overt evidence acquisition*, R observes $\hat{m}$ and π, updates his belief and

[7] The assumption $\hat{\omega} > x_0$ always makes equilibrium communication informative and can be interpreted as the conflict between the players' preferences being sufficiently large for a given G. Otherwise, if the conflict is small ($\hat{\omega} \leqslant x_0$), then, for some parameters of the model, equilibrium communication will be uninformative. In addition, it will always be uninformative if g is close enough to Dirac $\delta_{\hat{\omega}}$.

[8] For any compact metrizable Y, let ΔY denote the set of all Borel probability measures endowed with weak* topology.

[9] In principle, there can be many 'cheap-talk' messages that are always available to S. However, in this environment, any cheap-talk communication is uninformative because S's payoff is strictly increasing in R's posterior mean. Thus, it is without loss of generality to assume there is a unique 'cheap-talk' message $\hat{m} = \varnothing$ which can be interpreted as a S's claim that she does not have any proof.

chooses an action.

Note that in both variants of the game, R observes π if S discloses evidence. This assumption enables the 'hard evidence' interpretation of information. That is, if S discloses a piece of evidence certifying some statement about the state, such a certificate must also include a non-falsifiable description of the test that generated it.[10]

We refer to the probability ρ as the *reliability* of the testing environment and assume that it is fixed and commonly known. In many settings, this is motivated by the uncertainty about how long collecting evidence will take. For example, there might be a contracting deadline and the probability S is unable to obtain the evidence by the deadline might be independent of the chosen test.

There exist a number of interpretations of the payoff environment. First, as described above, ω can be interpreted as a single receiver's private information. Second, the set Ω can be viewed as a population of receivers. In this interpretation, S publicly discloses evidence and aims to maximize the mass of those who accept. Third, consider a setting in which R does not have a private type, but the action space is continuous. For example, suppose that R is taking an action $a \in A = [0,1]$ to match the state ($u_R(a,\theta) = -(a-\theta)^2$), and S has a state-independent utility function that is convex-concave in the action, i.e. $u_S = G$.[11] Then such a model is strategically equivalent to the one we study.[12]

We study perfect Bayesian equilibria of the game. However, because of the assumptions on the players' preferences, the analysis is amenable to the belief-based approach as explained below. Since it is straightforward to recover the players' actual strategies from beliefs, it will be convenient to abstract away from strategies in the main text of the paper.[13]

[10] An alternative but equivalent formulation of this conceptual assumption is that each test is a mapping $\pi\colon \Theta \to \Delta(M \times \Pi)$ such that each "extended message" (m', π') also encodes the description of the experiment, i.e. $\pi(M \times \{\pi\}|\theta) = 1$ for all θ and π. In this formulation, R would observe π only through the extended message in the event of disclosure.

[11] Dworczak and Martini (2019) provide an example of a continuous-action game in which the sender's objective is convex-concave.

[12] To see why, note that G measures S's indirect utility as a function of the induced posterior mean in either interpretation, the belief-based approach section below elaborates on this.

[13] Appendix A.1 presents a formal definition of an equilibrium.

Belief-based approach. We will follow a framework of representing information structures with convex functions which has proven convenient in information-design problems (Gentzkow and Kamenica, 2016; Kolotilin, 2018).

Fix any R's posterior belief $\beta \in \Delta\Theta$ with the mean $x^\beta := \int_\Theta \theta \, \mathrm{d}\beta(\theta)$. The best response of R with an outside option ω coincides with $a^\omega(\beta) := \mathbf{1}\{x^\beta \geqslant \omega\}$ for all $\omega \neq x^\beta$. Then, the S's interim expected payoff is given by the probability R with a posterior mean x^β accepts, i.e.

$$\int_\Omega u_S(a^\omega(\beta)) \, \mathrm{d}G(\omega) = G(x^\beta).$$

In other words, R's outside-option CDF G plays the role of S's *indirect utility function* defined on the set $X := [0,1]$ of *posterior means*.

Because both players' interim expected payoffs depend only on the mean of a posterior belief, each test π can be associated with a *posterior-mean distribution*, which we will identify with the corresponding CDF F_π.[14] Without loss of generality, because all relevant distributions are supported within $[0,1]$ and cannot have a mass at 0, we treat CDFs as functions over $[0,1]$. We will further identify each posterior-mean distribution with the corresponding *integral CDF* (ICDF), which is an increasing convex function I_π defined as the antiderivative of the CDF F_π[15]

$$\begin{aligned} I_\pi \colon [0,1] &\to [0,1], \\ x &\mapsto \int_0^x F_\pi. \end{aligned}$$

Clearly, the CDF can be recovered as the right derivative of the ICDF, $F_\pi = I'_\pi$.[16]

To describe the set of all feasible ICDFs, first note that the posterior-mean distribution of a fully-revealing test coincides with the prior $\overline{F}$ because each posterior is degenerate at the corresponding state. Second, the posterior-mean distribution $\underline{F}$ of any uninformative test has unit mass at the the prior mean x_0. Let $\overline{I}$ and $\underline{I}$ denote the ICDFs corresponding to full information and no information, respectively.

[14] That is, let $\beta \colon M \to \Delta\Theta$ be the belief map, i.e. any measurable map that satisfies the Bayes rule, $\int_{\hat{\Theta}} \pi(\hat{M}|\cdot) \, \mathrm{d}\overline{F} = \int_\Theta \int_{\hat{M}} \beta(\hat{\Theta}|\cdot) \, \mathrm{d}\pi(\cdot|\theta) \, \mathrm{d}\overline{F}(\theta)$ for all Borel $\hat{\Theta}, \hat{M} \subseteq [0,1]$. Then the posterior-mean CDF corresponding to β is given by $F_\pi(x) := \int_\Theta \pi(\{m \in M \colon x^{\beta(m)} \leqslant x\}|\cdot) \, \mathrm{d}\overline{F}$.

[15] We omit the variable of integration whenever it is unambiguous, using the standard notation: $\int_a^b f := \int_a^b f(x) \, \mathrm{d}x$, $\int_a^b f \mathrm{d}g := \int_a^b f(x) \, \mathrm{d}g(x)$.

[16] Throughout the paper, for any convex $I \colon [0,1] \to [0,1]$, let $I'(x)$ denote the right derivative of I at x for all $x \in [0,1)$ and $I'(1) := 1$.

Third, it is well known that the Blackwell informativeness order on information structures translates into mean-preserving spreads over distributions of posterior means.[17] Hence, we say that an ICDF J is *more informative* than a posterior-mean ICDF I (and I is less informative than J) if and only if $J(x) \geqslant I(x)$ for all $x \in [0,1]$ with equality at $x = 1$.

Because every test π is more (less) informative than an uninformative (fully informative) one, the corresponding posterior-mean ICDF I_π is a convex function satisfying $\bar{I} \geqslant I_\pi \geqslant \underline{I}$. Gentzkow and Kamenica (2016) and Kolotilin (2018) showed that the converse also holds, that is for any convex I such that $\bar{I} \geqslant I \geqslant \underline{I}$, there exists a test π such that $I_\pi = I$. Thus, we can define the set of all feasible posterior-mean ICDFs as

$$\mathcal{I} \coloneqq \{I \colon [0,1] \to [0,1], \text{ s.t. } I \text{is convex and } \bar{I} \geqslant I \geqslant \underline{I}\}.$$

Finally, in addition to the informativeness partial order $\geqslant$ on $\mathcal{I}$, we also define the strict informativeness order $>$ as an asymmetric part of $\geqslant$. That is, J is *strictly more informative* than I if and only if $J > I$, i.e., $J \geqslant I$ and $J \neq I$. In the current setting, this notion has the following interpretation: J is strictly more informative than I if and only if R is ex-ante strictly better off having posterior-mean ICDF J than I for any strictly increasing CDF of the outside option (see Corollary 6 in Appendix A.2).

3 Analysis

In this section, we characterize the equilibria of the game. We start by analyzing an auxiliary disclosure game in which the evidence structure is fixed and commonly known. Then, we characterize the resulting S value and show that the ex-ante acquisition problem can be stated as an optimization problem in the overt case and as a fixed-point problem in the covert case. Finally, we characterize the equilibrium evidence acquisition.

3.1 Voluntary disclosure

In this section, we briefly revisit an auxiliary Dye (1985) disclosure game in which the evidence structure I is fixed and commonly known. Analyzing this game is useful to understand the on-path R beliefs and S disclosure decisions. Moreover, in the overt case, I is

[17] Rothschild and Stiglitz (1970) show the equivalence in the context of a risk averter's preferences over monetary lotteries (see Leshno, Levy, and Spector, 1997, for a correction of the proof). Blackwell and Girshick (1954) prove a decision-theoretic equivalence result in the finite case.

always observed by R and so the auxiliary game can be treated as a subgame of the main game.

Fix any feasible posterior-mean ICDF $I \in \mathcal{I}$ with the corresponding CDF $F \coloneqq I'$. Which realizations of I should S disclose? Let $x_\varnothing \in X$ denote *non-disclosure R's posterior mean* and note that it is strictly optimal for S (not) to disclose x if and only if it is above (below) $x_\varnothing$ since her interim payoff function G is strictly increasing. Thus, in equilibrium, the best-responding disclosure threshold $d_{\rho,I}$ must coincide with $x_\varnothing$ which itself must be Bayes-consistent given the $d_{\rho,I}$-disclosure strategy, the reliability ρ, and F, which can be written as

$$d_{\rho,I} = x_\varnothing = \frac{(1-\rho)}{1-\rho+\rho F(d_{\rho,I})} x_0 + \frac{\rho F(d_{\rho,I})}{1-\rho+\rho F(d_{\rho,I})} \mathbb{E}_F(x | x \leqslant d_{\rho,I}). \tag{1}$$

The following lemma provides a convenient characterization of the solution of (1) using the ICDF approach.

Lemma 1. *In the Dye (1985) game with a fixed $I \in \mathcal{I}$, the disclosure threshold $d_{\rho,I}$ solves*

$$I(d_{\rho,I}) = \frac{1-\rho}{\rho}(x_0 - d_{\rho,I}). \tag{2}$$

Moreover,

(i) there is a unique solution to (2) in $\operatorname{co}(\operatorname{supp} I)$*;*[18]

(ii) more information leads to more disclosure: $d_{\rho,I}$ is strictly decreasing in ρ, and decreasing in I with respect to the informativeness order $\geqslant$;

(iii) perfect reliability leads to unraveling: $d_{1,I} = \min \operatorname{supp} I$, that is, S discloses all realizations of I (except, possibly, the lowest one).

The formal proof is omitted (all omitted proofs are presented in Appendix A.2) and the intuition is as follows. The equivalence of (1) and (2) is a straightforward application of the integration by parts in (1). Parts (i-iii) can be clearly seen from Figure 3. The uniquness holds because the equilibrium disclosure threshold must be at the intersection of the increasing (strictly on $[\min \operatorname{supp} I, 1]$) ICDF and the straight line whose negative slope (strictly if $\rho < 1$) depends on ρ.

[18] Moreover, there is a unique solution to (2) in $[0,1]$ if and only if $\rho \neq 1$ or $\min \operatorname{supp} I = 0$. In principle, R may have an off-path non-disclosure posterior mean $x_\varnothing < \min \operatorname{supp} I$. We will however, define $d_{1,I} \coloneqq \lim_{\rho \nearrow 1} d_{\rho,I} = \min \operatorname{supp} I$ for convenient continuity in ρ. While this is least permissive in terms of equilibria outcomes, it turns out to be without loss of generality.

At $\rho = 1$, the two lines intersect on $[0, \min \operatorname{supp} I]$ which means that S will disclose all (except, possibly, the lowest) realizations in $\operatorname{supp} I$. Intuitively, under full reliability, R is fully skeptical of non-disclosure which forces S to reveal everything. This can be seen as a special case of the unraveling principle (Milgrom, 1981; Grossman, 1981).

Next, note that a less informative I is pointwise higher, and a decrease in ρ makes the slope of the straight line steeper. Either of these two shifts leads to the intersection occuring at a higher posterior mean. Intuitively, when S is less informed, R's skepticism is more 'muted' which allows S to credibly conceal more evidence in equilibrium.[19]

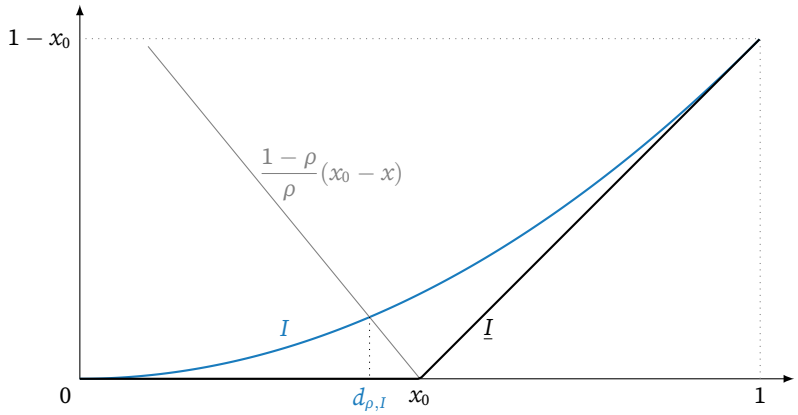


Figure 3: Construction of the disclosure threshold $d_{\rho,I}$.

3.2 Equilibrium Evidence Acquisition

In this section, we endogenize the evidence structure as S's ex-ante choice.

We begin with a few definitions. Say that I is an *(c-) o-equilibrium structure* if there exists a PBE (formally defined in Appendix A.1) of the (covert) overt acquisition game in which S chooses π such that $I_\pi = I$. Next, let $v_\rho(I|x_\varnothing)$ denote the S expected value assuming (i) some fixed R's non-disclosure belief mean $x_\varnothing \in X$, (ii) some chosen evidence structure $I \in \mathcal{I}$, and (iii) S discloses realizations $x > x_\varnothing$. Formally,

$$v_\rho(I|x_\varnothing) := \left[1 - \rho + \rho I'(x_\varnothing)\right] G(x_\varnothing) + \rho \int_{x_\varnothing}^{1} G \,\mathrm{d}I' - G(x_0).$$

[19] Similar uniqueness and comparative statics results were established in Propositions 1 and 2 in Jung and Kwon (1988) (see also Proposition 1 in Acharya, DeMarzo, and Kremer, 2011) for continuous distributions and in Corollary 2 and Proposition 2 of DeMarzo et al. (2019) for general distributions.

Note that by subtracting the no-information payoff $G(x_0)$, we normalize $v_\rho(\underline{I}|x_\varnothing)$ to zero for all $x_\varnothing \in X$ and $\rho \in (0,1]$. This definition enables the following preliminary characterization of equilibria.

Lemma 2. *For any evidence structure $I^* \in \mathcal{I}$,*

(i) I^* is an o-equilibrium structure if and only if

$$I^* \in \operatorname*{argmax}_{I \in \mathcal{I}} v_\rho(I|d_{\rho,I}), \tag{Overt}$$

(ii) I^* is an c-equilibrium structure if and only if

$$I^* \in \operatorname*{argmax}_{I \in \mathcal{I}} v_\rho(I|d_{\rho,I^*}), \tag{Covert}$$

Notice an important difference between the two seemingly similar programs: while (Overt) is an optimization problem, (Covert) is a fixed-point problem. Conceptually, in the overt case, S can commit to the way information is acquired (subject to reliability) but not to the way it is disclosed. That is, a deviation to any I will lead to a Bayes-consistent R's non-disclosure posterior mean $d_{\rho,I}$. Then, since the disclosure subgame for each I has a unique outcome, S will choose the best such outcome across all feasible evidence structures.

In contrast, in the covert case, a deviation to some evidence structure I is not detected by R and so his non-disclosure posterior remains the same as on the equilibrium path. Therefore, S chosen evidence structure I^* must be best-responding to a fixed non-disclosure posterior mean $x_\varnothing$ which itself must be Bayes-consistent with I^*, that is, $x_\varnothing = d_{\rho,I^*}$.

The above lemma has two immediate corollaries. First, by standard arguments using Weierstrass Theorem and Kakutani-Glicksberg-Fan Theorem, respectively, we have existence of o- and c-equilibria.

Corollary 1. *For any $\rho \in (0,1]$, an o-equilibrium and a c-equilibrium exist.*

Second, the o-equilibrium S payoff is unique in the overt case and is weakly above that any c-equilibrium S payoff. This is because S has a Stackelberg-like first mover advantage in the overt case. Therefore, she cannot do worse in an o-equilibrium than choosing any c-equilibrium structure.

It will also be useful to relate the (Overt) objective to the S full-commitment problem

in which she is able to directly design R's information. Let

$$\begin{aligned} v\colon \mathcal{I} &\to \mathbb{R}, \\ I &\mapsto v_1(I|0) = \int_0^1 G \,\mathrm{d}I' - G(x_0), \end{aligned}$$

denote the S indirect value function over R's posterior-mean ICDFs. Then the S full-commitment problem can be written as

$$\max_{I\in\mathcal{I}} \int_0^1 G \,\mathrm{d}I' = \max_{\mathcal{I}} v. \tag{FC}$$

Next, for any given $I \in \mathcal{I}$, let $I_\rho^D \in \mathcal{I}$ denote the *disclosed* evidence structure, that is, the distribution of R's posteriors. Since S does not disclose either when she is uninformed or when the realized evidence is below $d_{\rho,I}$ and otherwise R's posterior mean equals exactly the realized evidence, a direct computation yields the *disclosed CDF*

$$I_\rho^{D\prime}(x) = \begin{cases} 0, & x < d_{\rho,I} \\ 1-\rho+\rho F(x), & x \geqslant d_{\rho,I} \end{cases}$$

which gives the following simple expression for the *disclosed ICDF*

$$I_\rho^D(x) \coloneqq \left[\rho I(x) + (1-\rho)(x - x_0)\right]^+, \tag{Disclosed}$$

where $[z]^+ \coloneqq \max(z, 0)$.

In this case, Lemma 1 and (Disclosed) imply that the (Overt) objective evaluated at some ICDF can be written as the (FC) objective evaluated at the disclosed ICDF, that is

$$\begin{aligned} v_\rho(I|d_{\rho,I}) &= \left[1-\rho+\rho I'(d_{\rho,I})\right] G(d_{\rho,I}) + \rho \int_{d_{\rho,I}}^1 G \,\mathrm{d}I' - G(x_0) \\ &= \int_0^1 G(x) \,\mathrm{d}\left[\rho I(x) + (1-\rho)(x-x_0)\right]^{+\prime} - G(x_0) \\ &= v(I_\rho^D) \end{aligned}$$

3.3 Benchmark: Perfect Reliability

Before we characterize the equilibrium evidence structure, it will be instructive to look at the extreme case of $\rho = 1$, that is, when S always obtains the evidence she seeks. Recall that in this case, a standard unraveling argument applies (Lemma 1 part (iii)), that is, S fully discloses all evidence she obtains due to R being fully skeptical, hence, $I_1^D = I$.

Then, both (Overt) and (Covert) programs reduce to (FC). This means that two out of three forces affecting R's information—voluntary disclosure and observability of acquisition strategy—are irrelevant in this case and an equilibrium is characterized by a pure information design problem. Kolotilin et al. (2017), Kolotilin (2018) study a model of Bayesian persuasion with R's private payoff type which is similar to the above. In particular, their results imply that if the distribution of R types is unimodal, the optimal signal is a *t upper censorship* – it reveals (pools) all states below (above) some threshold $t \in \Theta$.[20]

In other words, one can solve the overt and covert cases *under full reliability* using a well-known methods from information design. Below we explicitly state a useful lemma (based on the ICDF approach of Lipnowski, Ravid, and Shishkin (2021)) which not only delivers the (FC) optimality of upper censorships, but also turns out to be able to significantly simplify the analysis of the imperfect reliability case as we will see in the next sections.

Intuitively, when the distribution of outside options is unimodal, the S indirect value function G is convex below and concave above $\hat{\omega}$. Therefore, when the state is low (high), more information benefits (hurts) S. To formalize this intuition, it will be useful to rewrite the objective function v by integrating by parts twice as follows

$$v(I) = \int_0^1 G \,\mathrm{d}I' - G(x_0) = \int_0^1 G \,\mathrm{d}(I' - \underline{I}') = \int_0^1 (I - \underline{I}) \,\mathrm{d}g.$$

Such an integral representation implies that the S's value can be visualized (see Figure 4a) as the 'area' between I and $\underline{I}$ 'weighted' by the density g of R's outside option and can be decomposed into the positive part $\int_0^{\hat{\omega}} (I - \underline{I}) \,\mathrm{d}g$ and the negative part $\int_{\hat{\omega}}^1 (I - \underline{I}) \,\mathrm{d}g$.

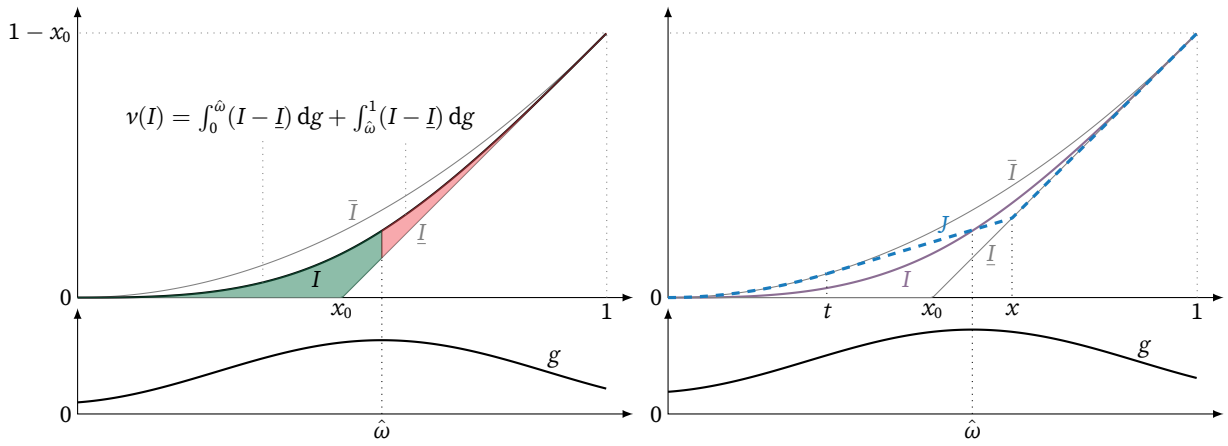
This decomposition motivates the following definitions. Call J a *pivoted* I if J is weakly above I on $[0, \hat{\omega}]$ and weakly below I on $[\hat{\omega}, 1]$ and $I \neq J$. Call J an *S-improvement over* I if $v(J) - v(I) = \int (J - I) \,\mathrm{d}g > 0$ for all strictly quasiconcave g with a peak at $\hat{\omega}$. The next lemma shows that these two relations coincide.[21]

Lemma 3. *For any $I, J \in \mathcal{I}$, J is an S-improvement over I if and only if J is a pivoted I.*

Next, we establish that only upper censorships are immune to S-improvements which

[20] Optimality of upper censorship in similar settings also appears in Alonso and Câmara (2016b) and Dworczak and Martini (2019).

[21] Note that if $\hat{\omega} = 1$, then S-improvements correspond to strictly more informative structures, and pivoting correspond to a mean-preserving spread. Hence, Lemma 3 can be seen as a generalization of the strict version of the Blackwell-Rothschild-Stiglitz equivalence result (see Corollary 6 in Appendix A.2).



(a) $v(I)$ as the sum of the positive and negative parts.

(b) Lemma 4: J is an S-improvement over I.

Figure 4: Decomposition of v and illustration of an S-improvement in Lemma 4.

follows from a construction in Lipnowski et al. (2021) showing how any I can be pivoted to obtain an upper censorship J as a result.

Lemma 4 (Lipnowski et al. (2021), Lemma 5). *For any $I \in \mathcal{I}$, there exists some upper censorship which is either an S-improvement over I or coincides with it.*

The construction is illustrated in Figure 4b. Fix any $I \in \mathcal{I}$. By Lemma 3, we need to construct a t upper censorship J which either coincides with I or is a pivoted I. Take the line tangent to $\bar{I}$ going through the point $(\hat{\omega}, I(\hat{\omega}))$, and let $(t, \bar{I})$ and $(x, \underline{I})$ be the points of tangency with $\bar{I}$ and intersection with $\underline{I}$, respectively. Let J be equal to the tangent line on $[t, x]$, to $\bar{I}$ on $[0, t]$, and to $\underline{I}$ on $[x, 1]$. It is easy to verify that J is a t upper censorship and that either $I \neq J$ and then J is an S-improvement, or $I = J$, then no improvement exists.

Since any non-upper-censorship can be S-improved and v is simultaneously the (FC) objective, and (Overt) and (Covert) objective under $\rho = 1$, we immediately obtain the following characterization of the perfect reliability case.

Corollary 2. *If $\rho = 1$, then there exists some $t_1^* \in (0, \hat{\omega})$ such that the t_1^* upper censorship is the unique solution of the* (FC)*,* (Overt)*, and* (Covert) *problems.*

3.4 Overt Acquisition

Now we turn to the general overt case with imperfectly reliable testing. We start by introducing the following class of information structures that nests the upper censorship

defined in Section 3.3.

Call an evidence structure $I \in \mathcal{I}$ *a* (θ_l, θ_h) *two-sided censorship of* $J \in \mathcal{I}$ if it is a garbling of J which perfectly reveals all realizations in $[\theta_l, \theta_h]$, pools the ones above θ_h, and also pools the ones below θ_l. Formally, I is the lowest ICDF that coincides with J on $[\theta_l, \theta_h]$ as illustrated in Figure 5. This class includes three important special cases: I is *a* θ_h *upper censorship* of J if $\theta_l = 0$; *a* θ_l *lower censorship* of J if $\theta_h = 1$; and *a* θ *pass/fail test* of J if $\theta_l = \theta_h = \theta$. Whenever $J = \bar{I}$ in the above notions, we will omit saying 'of $\bar{I}$' for brevity.

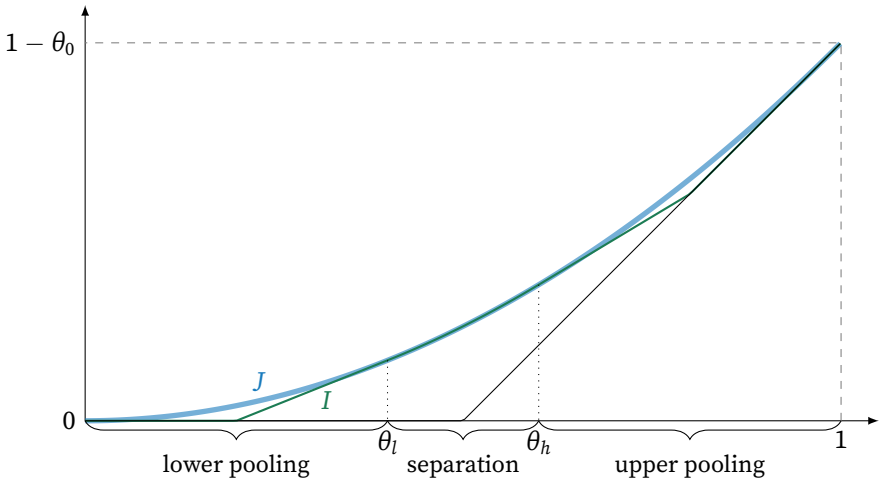


Figure 5: Two-sided censorship.

A test inducing a two-sided censorship can be interpreted as a grading system that assigns the PASS grade to the states above the upper cutoff, the FAIL grade to the states below the lower cutoff, and has a variety of intermediate grades corresponding exactly to each state in between. In addition, if $\theta_l = 0$ and $\theta_h = 1$, both pooling intervals are empty, which corresponds to the fully informative structure $\bar{I}$. And if $\theta_l = \theta_h \in \{0, 1\}$, then all states are pooled, which corresponds to the uninformative structure $\underline{I}$.

In order to state the main results, we now introduce a notion of disclosure-equivalence to address the multiplicity of equilibrium evidence structures that naturally arises in the model.

Definition 1. Call I and J *disclosure-equivalent* if their (Disclosed) transforms coincide, that is, $I^D_\rho = J^D_\rho$.

To illustrate this definition, suppose I is an equilibrium evidence structure induced by

a test which is perfectly informative about states below some $x \leqslant d_{\rho,I}$. Then, although S obtains precise information about states below x, she will end up not disclosing any of the corresponding realizations of I. Now consider J which is a x lower censorship of I, that is, J pools all realizations of I below x. But then this is observationally equivalent from R's perspective since the same realizations of I and J are disclosed and so it does not matter whether S learns more or less bad news which will be concealed anyway. As a result, disclosure equivalence affects neither the Bayes-consistent non-disclosure posterior, nor whether (MP) is satisfied, nor whether (Overt) and (Covert) equilibrium conditions are satisfied.

It is also easy to verify that Definition 1 and (Disclosed) imply that I and J are disclosure equivalent if and only if they coincide on $[d_{\rho,I}, 1]$. This implies that the disclosure-equivalence class of any I has the least informative element given by the $d_{\rho,I}$ lower censorship of I. From now on, we will focus on such least informative equilibria structures. The reason for such a selection from disclosure-equivalence classes is three-fold. First, it is straightforward to construct a disclosure-equivalence class from the least informative structure and thereby recover all equilibria.[22] Second, this selection can be seen as a 'revelation principle': for every equilibrium of the game, there exists a 'canonical' outcome-equivalent equilibrium, in which there is a unique bad-news realization which is the only one not disclosed by S. Third, one can also view this as a selection based on vanishing Blackwell-monotone cost of acquiring information.

The following theorem provides a characterization of o-equilibria.

Theorem 1. *There exists $\overline{\rho}^o \in [0,1)$ such that any o-equilibrium evidence structure is disclosure equivalent to a pass/fail test if $\rho < \overline{\rho}^o$, and to the $(d_{\rho,\bar{I}}, t_1^*)$ two-sided censorship if $\rho > \overline{\rho}^o$.*

Moreover, for all except countably many $\rho < \overline{\rho}^o$, the equilibrium pass/fail threshold $t_\rho^o \in (0, d_{\overline{\rho}^o,\bar{I}})$ is unique and strictly increasing in ρ.

Recall that in isolation, the voluntary disclosure force leads to pooling at the bottom and the information design force leads to pooling at the top of the state space as evident from Sections 3.1 and 3.3, respectively. Theorem 1 then demonstrates that whether and how these two forces interact depends on reliability. When reliability is above $\overline{\rho}^o$, the interaction between the two forces is trivial and optimal evidence structure is a two-sided censor-

[22] Indeed, the set of all structures disclosure-equivalent to I is the $\geqslant$-interval between the $d_{\rho,I}$ lower-censorship of I and the pointwise maximum over all structures coinciding with I on $[d_{\rho,I}, 1]$.

ship of the state. The lower threshold $d_{\rho,\bar{I}}$ is not affected by the design of the evidence structure and coincides with the disclosure threshold under fully-revealing evidence structure. Moreover, the upper threshold t_1^* is unaffected by voluntary disclosure: it stays constant and coincides with the optimal upper threshold that the sender would use under $\rho = 1$. In other words, the optimal structure is a straightforward combination of the two forces.

However, when reliability is below $\overline{\rho}^o$, the interaction between the two forces becomes non-trivial and the sender switches to a pass/fail test. From the ex-ante perspective, S prefers more information at the bottom and, therefore, voluntary disclosure hurts S because it induces pooling of low states. In other words, while she would want to commit to reveal low states, she cannot if disclosure is voluntary. When ρ drops below $\overline{\rho}^o$, it becomes optimal to design evidence structure in order to reduce the ex-ante loss from lower pooling. This is achieved by a pass/fail test, as it allows to reduce the lower pooling interval by enlarging the upper pooling interval because, under pass/fail test, S discloses if and only if she passes the test.

Moreover, as reliability falls, so does the total probability of disclosure, if the signal is kept the same. Then, it is optimal to lower the pass/fail test threshold in order to enlarge the upper pooling interval and increase the probability of disclosure conditional on obtaining evidence thereby compensating for falling total probability of disclosure.

Formally, the result is based on two observations.

Corollary 3 (of Lemma 4)**.** *Every o-equilibrium structure is disclosure-equivalent to an upper censorship.*

Proof. Take any o-equilibrium structure I. By Lemma 4, there exists an S-improvement upper-censorship J. Then, $J - I$ is non-negative on $[0, \hat{\omega}]$ and non-positive on $[\hat{\omega}, 1]$ and so is $J_\rho^D - I_\rho^D$. If $J_\rho^D = I_\rho^D$, then I is disclosure equivalent to an upper censorship. Otherwise, J_ρ^D is an S-improvement over I_ρ^D and therefore I does not maximize $v(I_\rho^D) = v_\rho(I|d_{\rho,I})$ which contradicts with it being an o-equilibrium structure, by Lemma 2. □

This observation suggests that the information design force has a similar effect as we observed in the case of perfect reliability in Section 3.3. It allows to relax the (Overt) program to a one-dimensional optimization with respect to upper censorship thresholds $t \in \Theta$. But then every upper censorship is disclosure-equivalent to either a two-sided censorship or to a pass/fail test depending on whether the optimal upper censorship threshold t_ρ^o is

above or below the corresponding non-disclosure posterior.

Second, we demonstrate that the relaxed optimization objective has the increasing differences property with respect to t and ρ so that the set of optima is increasing in ρ in the sense of strong set order. Moreover, it has strictly increasing marginal differences for low values of t which implies that the set of optima is either equal to $\{t_1^*\}$ or weakly below t_1^* it with every selection strictly increasing in ρ. Denoting by $\overline{\rho}^o$ the switching point, this implies that the o-equilibrium upper censorship threshold is unique for all $\rho > \overline{\rho}^o$ and all except possibly a countable subset of $\rho \leqslant \overline{\rho}^o$.

Remark. *Figure 2a illustrates how o-equilibrium changes with reliability in the case for a particular numeric example where the reliability cutoff $\overline{\rho}^o$ satisfies the equation $t_1^* = d_{\overline{\rho}^o,\bar{I}}$. That is, the switching happens when the lower pooling region becomes just large enough to intersect the upper pooling region of the optimal two-sided censorship. While this is often true, Theorem 1 only guarantees that $t_1^* \geqslant d_{\overline{\rho}^o,\bar{I}}$. That is, the proof relies on the optimization comparative statics results that cannot rule out the possibility that the solution might have a discontinuity with respect to the parameter ρ in general.*

3.5 Covert Acquisition

We now turn to the case in which the S acquisition strategy is unobserved by R unless S discloses evidence. The following result shows that despite the additional ex-ante S incentive constraint, c-equilibria structures share some qualitative properties of o-equilibria. At the same time, the comparative statics result with respect to reliability is reversed.

Theorem 2. *There exists $\overline{\rho}^c \in [0, \rho^o]$ such that every c-equilibrium evidence structure is disclosure equivalent to a pass/fail test if $\rho \leqslant \overline{\rho}^c$, and to the $(d_{\rho,\bar{I}}, t_1^*)$ two-sided censorship if $\rho > \overline{\rho}^c$.*

Moreover, for all $\rho < \overline{\rho}^c$, the c-equilibrium pass/fail threshold $t_\rho^c \in (d_{\overline{\rho}^c,\bar{I}}, x_0)$ is unique and strictly decreasing in ρ, and $t_{\overline{\rho}^c}^c = d_{\overline{\rho}^c,\bar{I}} = t_1^*$.

To explain the intuition behind the result, it will be useful to establish some preliminary observations. First, we make the following difference-in-difference comparison between overt and covert deviations.

Lemma 5. *The net benefit to S from an ex-ante deviation from I^* to I in the covert case is higher (lower) than that in the overt case if and only if d_{ρ,I^*} is higher (lower) than $d_{\rho,I}$.*

Proof. The difference in differences equals[23]

$$\begin{aligned}&[\nu_\rho(I|d_{\rho,I^*}) - \nu_\rho(I^*|d_{\rho,I^*})] - [\nu_\rho(I|d_{\rho,I}) - \nu_\rho(I^*|d_{\rho,I^*})]\\ &= (1-\rho)\left[G(d_{\rho,I^*}) - G(d_{\rho,I})\right] + \rho\int_{d_{\rho,I}}^{d_{\rho,I^*}} I'\,\mathrm{dg},\end{aligned}$$

and, thus, has the same sign as $d_{\rho,I^*} - d_{\rho,I}$. □

Intuitively, this is because the only way S may benefit from covertly deviating to I *in addition* to her gain from an overt deviation to the same I is by not disclosing its realization and obtaining the payoff corresponding to a higher on-path posterior mean than the one Bayes-consistent with the deviation.

This logic has a tight connection to the so-called "minimum principle" of DeMarzo et al. (2019). In our language, that paper studies a constrained covert evidence acquisition game with a uniform distribution of outside options. Their results imply that c-equilibria in the uniform case are characterized by the *minimum principle* which can be stated in our setting as

$$I^* \in \operatorname*{argmin}_{I\in\mathcal{I}} d_{\rho,I}. \tag{MP}$$

With linear indirect payoff function G, the information-design incentives are absent, or equivalently any distribution of R posterior means has the same ex-ante value for S.[24] Thus, Lemma 5 would imply that I^* is a c-equilibrium structure if and only there is no other I with a lower Bayes-consistent non-disclosure posterior mean $d_{\rho,I}$, or equivalently that I^* is minimal in the sense of (MP).

In our non-uniform case, the minimum principle is no longer necessary in the non-uniform case, but it turns out that the combination of (MP) and (Overt) optimality is sufficient to solve the (Covert) program.

Corollary 4 (of Lemma 5)**.** *If the set of o-equilibria satisfying the minimum principle* (MP) *is non-empty, then this set coincides with the set of c-equilibria.*

Now notice that with high reliability the unique o-equilibrium satisfies the minimum principle. As Figure 3 illustrates, the more information (higher ICDF) in the vicinity of the

[23] For any $a, b \in [0,1]$, we follow the notational convention $\int_a^b F\,\mathrm{dg} \coloneqq \int_0^b F\,\mathrm{dg} - \int_0^a F\,\mathrm{dg}$.

[24] Indeed, with $G(x) = x, x \in [0,1]$, *any* $I \in \mathcal{I}$ solves both the (Overt) and (FC) problems because $\nu_\rho(I|d_{\rho,I}) = \int_0^1 x\,\mathrm{d}I' - x_0 \equiv 0$.

non-disclosure posterior, the lower non-disclosure posterior is (Figure 3). The o-equilibrium two-sided censorship is disclosure equivalent to the t_1^* upper-censorship which is perfectly informative about low states and, therefore, minimizes the non-disclosure posterior. By Lemma 5, a covert deviation from such a structure cannot be more profitalble than an overt deviation. Because the latter is not profitable in an o-equilibrium, any o-equilibrium is a c-equilibrium and vice versa. In other words, with high reliability, not only information design and voluntary disclosure do not interact, but also the covertness of acquisition has no impact because S chooses a relatively detailed test.

In contrast, with low reliability, the o-equilibrium pass/fail test fails the minimum principle, because the threshold is always strictly below the minimal non-disclosure posterior $d_{\rho,\bar{I}}$. So there is an additional benefit to S from deviating from such a signal to some structure with a lower corresponding non-disclosure posterior which explains why c-equilibria may differ from o-equilibria. Still, a combination of Lemmas 4 and 5 allows to quickly establish that every c-equilibrium structure is disclosure-equivalent to an upper censorship. Namely, similarly to the argument in Corollary 3, if I is a c-equilibrium which is not disclosure-equivalent to its S-improvement upper censorship J, then J_ρ^D is an S-improvement over I_ρ^D. Hence, by Lemma 5, the benefit from deviating from I to J is strictly positive, because it would be a strict improvement in the overt case and J has a lower corresponding non-disclosure posterior.

However, to establish the comparative statics result of Theorem 2, the above observation is insufficient because it does not bear any implications off the equilibrium path. It turns out that a modification of Lemma 4 can be used to establish the dominance of the upper censorship class even off-path in the following sense. Then, solving the (Covert) program is equivalent to finding a pair $(I, x_\varnothing)$ such that S is best-responding to R's non-disclosure belief (i.e., I maximizes $v_\rho(\cdot|x_\varnothing)$ over $\mathcal{I}$), and that the R's belief is Bayes consistent (i.e., $x_\varnothing = d_{\rho,I}$). The following observation implies that every best-response is payoff-equivalent to an upper censorship.

Corollary 5 (of Lemma 4). *For all $x_\varnothing \in [0, x_0]$, every maximizer of $v_\rho(\cdot|x_\varnothing)$ over $\mathcal{I}$ coincides with some upper-censorship on $[x_\varnothing, 1]$, and the set of maximizers is independent of ρ.*

Proof. For any $x_\varnothing \in [0, x_0]$, rewrite the objective $v_\rho(I|x_\varnothing)$ can be rewritten as

$$\begin{aligned}\operatorname*{argmax}_{I\in\mathcal{I}} v_\rho(I|x_\varnothing) &= \operatorname*{argmax}_{I\in\mathcal{I}} \left[1-\rho+\rho I'(x_\varnothing)\right] G(x_\varnothing) + \rho\int_{x_\varnothing}^1 G\,\mathrm{d}I' - G(x_0)\\ &= \operatorname*{argmax}_{I\in\mathcal{I}} \int_0^1 G_{\vee x_\varnothing}\,\mathrm{d}(I'-\underline{I}')\\ &= \operatorname*{argmax}_{I\in\mathcal{I}} \int_0^1 (I_{\vee x_\varnothing}-\underline{I})\,\mathrm{d}g\end{aligned}$$

where we define $J_{\vee x_\varnothing}(x) \coloneqq \max\{J(x), J(x_\varnothing)\}$ for all $x \in X, J \in \mathcal{I}$ and $\mathcal{I}_{\vee x_\varnothing} \coloneqq \{J_{\vee x_\varnothing} : J \in \mathcal{I}\}$. It follows immediately that the set of maximizers is independent of ρ. Then, the definition of the function v and the notion of S-improvement can be readily extended to $\mathcal{I}_{\vee x_\varnothing}$. The rest of the argument is very similar to the proof of Corollary 3.

Take any solution I of the above program and consider its upper censorship S-improvement J as given by Lemma 4. Then, $J-I$ is nonnegative on $[0,\hat{\omega}]$ and non-positive on $[\hat{\omega},1]$ and so is $J_{\vee x_\varnothing} - I_{\vee x_\varnothing}$. If $J_{\vee x_\varnothing} = I_{\vee x_\varnothing}$, then I and J coincide on $[x_\varnothing, 1]$ and we are done. Otherwise, $J_{\vee x_\varnothing} \neq I_{\vee x_\varnothing}$ and so $J_{\vee x_\varnothing}$ is an S-improvement over $I_{\vee x_\varnothing}$, hence $v(J_{\vee x_\varnothing}) > v(I_{\vee x_\varnothing})$, which contradicts with I solving the program. □

In other words, when S is choosing a test under some fixed non-disclosure belief $x_\varnothing$, she can always guarantee herself a payoff of $G(x_\varnothing)$ by non-disclosing. Hence, best-responding to $x_\varnothing$ is equivalent to solving the full-commitment problem with the ex-post payoff function $G_{\vee x_\varnothing}$ which is equal to G truncated from below at $G(x_\varnothing)$. Therefore, Lemma 4 implies that pivoting the ICDF on $[x_\varnothing, 1]$ constitutes an S-improvement to which only upper censorship structures are immune to.

Next, the (Covert) problem reduces to finding a pair $(t, x_\varnothing)$ such that $x_\varnothing = d_{\rho,I_t}$ and the t upper censorship I_t is a best response to $x_\varnothing$, that is t maximizes $\int_0^1 G_{\vee x_\varnothing}\,\mathrm{d}I'_t$ over Θ. Figure 7 illustrates the graphs of the best-response and the Bayes-consistency mappings and their intersections for various reliability levels. It is easy to show that the S best response is unique, continuous in $x_\varnothing$, equals t_1^* for $x_\varnothing \leqslant t_1^*$ and strictly increasing otherwise. But then since the Bayes consistency mapping $t \mapsto d_{\rho,I_t}$ is continuous, equal to $d_{\rho,\bar{I}}$ below the diagonal and strictly decreasing above. The intersection of the graphs of the best-response and Bayes-consistency mappings then exists and unique. Moreover, since higher reliability does not change the best response and lowers the Bayes consistent non-disclosure belief, the intersection is constant when $d_{\rho,\bar{I}} \leqslant t_1^*$ and strictly decreasing otherwise. In other words, the key reason behind the comparative statics of the c-equilibrium

pass/fail threshold is the skepticism effect uncovered in Lemma 1. Finally, one can define $\overline{\rho}^c := \inf\{\rho \in (0,1] \colon d_{\rho,\bar{I}} \leqslant t_1^*\}$ to obtain the result.

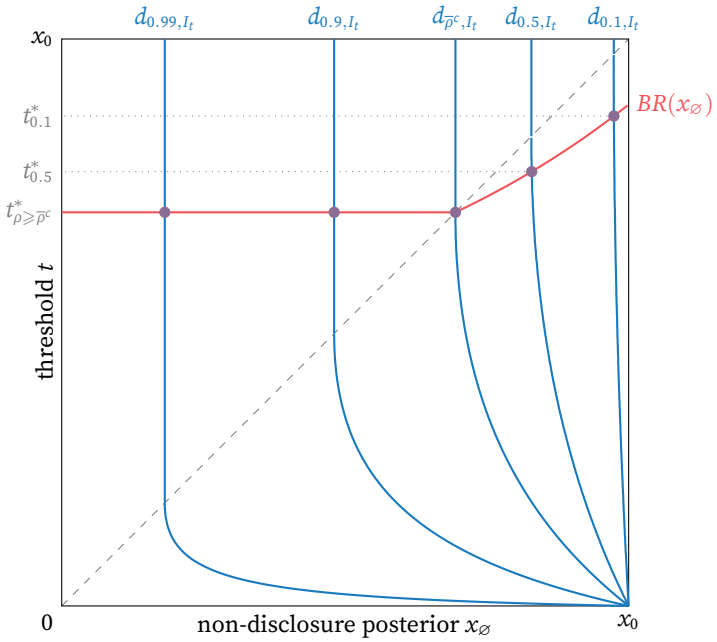


Figure 6: Overt acquisition

Figure 7: C-equilibrium thresholds obtained for various reliability levels as intersections of the best-response and Bayes-consistency mappings for the case of uniform θ and triangularly distributed ω with the peak at $3/5$.

3.6 Covert vs overt acquisition

In this section, we compare each player's equilibrium payoffs between the overt and covert cases.

First, note that in the case of high reliability ($\rho > \overline{\rho}^o$), all o-equilibrium and c-equilibrium structures are disclosure equivalent to the same structure. Therefore, each player would be indifferent between the overt and covert cases.

Second, note that it follows immediately from Lemma 2 that for any ρ, the sender's payoff is always weakly higher in overt case than in the covert case. Intuitively, compared to the (Covert) program, in the (Overt) program, the sender has a Stackelberg-like first-mover advantage, or, equivalently, any c-equilibrium structure is feasible in (Overt) program.

Finally, we turn to the comparison of the receiver's payoffs from c- and o-equilibrium pass/fail tests. The following proposition shows that the receiver prefers the sender's evidence acquisition to be covert rather than overt when reliability is low.

Proposition 1. *For any reliability level $\rho \in (0, \overline{\rho}^c)$, the receiver's equilibrium payoff is strictly higher under covert acquisition than under overt acquisition.*

The intution is as follows. First, we note that Theorem 1 and Theorem 2 imply that the c-equilibrium pass/fail threshold is always strictly above than the o-equilibrium one (see Figure 2 for an illustration). Second, we can show that the receiver's payoff from the information structure $[I_t]^D_\rho$ corresponding to the (Disclosed) t-threshold pass/fail test is strictly increasing in the threshold t for $t \in [0, d_{\rho,\bar{I}}]$. In fact, this follows from a much stronger observation: while the pass/fail structures I_t with different thresholds are never Blackwell-ranked, the sender-disclosed structure $[I_t]^D_\rho$ is strictly Blackwell-increasing for $t \in [0, d_{\rho,\bar{I}}]$. Combining these observations, we obtain that the receiver always strictly prefers the equilibrium outcome in the covert case to that in the overt case.

4 Conclusion

This paper studies overt and covert acquisition of hard information subject to imperfect reliabilty. We show how the tools from information design can be adapted to fully characterize the equilibrium evidence without putting parametric restrictions on a rich environment, despite the fact that the sender lacks full commitment. The main results demonstrate how each of the main forces—information design, voluntary disclosure, and covert/overt nature of acquisition—contribute to the equilibrium structure. When the reliability is high, the three forces do not interact: the sender acquires essentially the same signal (upper censorship) as under full commitment and the nature of acquisition is irrelevant. When the reliability is low, the equilibrium signal takes a very simple form of a pass/fail test with the threshold jointly determined by the three forces. In particular, the pass/fail threshold is monotone in reliability but whether it is increasing or decreasing depends on whether acquisition is overt or covert.

Our analysis under the assumptions of costless acquisition and exogenous reliability may also shed some light at situations when these assumptions fail to hold. First, if acquiring a test comes at some Blackwell-monotone cost, our results suggest that in some cases it

may have little impact on the equilibrium structure. In particular, when reliability is low, if a pass/fail test—the coarsest informative structure—arises in the costless case, then the sender might be even less likely to choose a more informative and complex structure at a positive cost. Second, suppose the sender could make an investment in reliability. Then our results can be seen as deriving the value of reliability which can then be compared to the cost of investment.

Alternatively, suppose, in the overt case, the sender could jointly choose among tests with various reliability levels which are all below some technological limit $\rho^{\max}$. Then, in the overt case, it could be easily shown[25] that the sender always strictly benefits from higher reliability and so she would prefer tests with reliability equal to $\rho^{\max}$ which could then be interpreted the exogenous reliability in our model.

More generally, suppose $\rho^{\max}$ might be test-specific under overt acquisition. For example, assume that more informative tests have lower $\rho^{\max}$. Then, our results suggest that the sender will have an additional incentive to coarsen the signal on top of the effect discussed in our main model. In particular, if all binary tests had the same $\rho^{\max}$, then the optimality of the pass/fail test will be preserved because deviating to a more informative test would be even less profitable.

□

[25] This holds due to replicability: any disclosed information structure that is feasible under lower reliability is also feasible under higher reliability. Formally, using (Disclosed), rewrite the (Overt) problem as maximizing v over $\{I^D_\rho \colon I \in \mathcal{I}\}$ and note that the constraint set is monotone in ρ with respect to set inclusion. The replicability principle does not hold in the covert case and it is possible that the sender may actually prefer lower reliability due to equilibrium effects.

A Appendix

A.1 Equilibrium Definition

Let Π be the set of all tests (i.e., measurable mappings $\Theta \to \Delta M$) endowed with the discrete σ-algebra. For any convex measurable space Y, given a probability measure $\nu \in \Delta Y$, let $\mathbf{E}\nu \coloneqq \int_Y y \,\mathrm{d}\nu(y) \in Y$ denote the barycenter of ν.

Under both overt and covert acquisition, an equilibrium consists of four objects: an S testing strategy $\pi \in \Pi$, an S's disclosure strategy (in terms of the probability of disclosure) $\delta\colon M \times \Pi \to [0,1]$, a R's belief map $\beta\colon (M \cup \{\varnothing\}) \times \Pi \to \Delta\Theta$, and a R's acceptance strategy (in terms of the probability of acceptance) $\alpha\colon (M \cup \{\varnothing\}) \times \Omega \times \Pi \to [0,1]$. For convenience, for all $\pi \in \Pi$ and $\omega \in \Omega$, denote $\delta_\pi \coloneqq \delta(\cdot,\pi), \beta_\pi \coloneqq \beta(\cdot,\pi), \alpha_\pi^\omega \coloneqq \alpha(\cdot,\omega,\pi), u_R^\omega \coloneqq u_R(\cdot,\omega)$ and let $\ell_{\pi,\rho,\delta}\colon \Theta \to \Delta(M \cup \{\varnothing\})$ denote the conditional likelihood (of messages) function corresponding to the experiment π with reliability ρ, i.e., for all Borel $M' \subseteq M$, $\ell_{\pi,\rho,\delta}(M'|\theta) \coloneqq \rho \int_{M'} \delta_\pi \,\mathrm{d}\pi(\cdot|\theta)$.

Now, an overt-acquisition equilibrium, or *o-equilibrium*, is a tuple $(\pi^*, \delta, \alpha, \beta)$ of measurable mappings such that, for all $m \in M, \omega \in \Omega, \pi \in \Pi$,

$$\beta_\pi \text{ is derived from Bayes rule given } \mu_0 \text{ and } \ell_{\pi,\rho,\delta}. \tag{o-Bayes}$$

$$\operatorname{supp} \alpha_\pi^\omega(m) \subset \operatorname*{argmax}_{a \in [0,1]} \int_\Theta u_R^\omega(a,\theta) \,\mathrm{d}\beta_\pi(\theta|m), \tag{R-IC}$$

$$\delta_\pi(m) \in \operatorname*{argmax}_{d \in [0,1]} \int_\Omega (d\alpha_\pi^\omega(m) + (1-d)\alpha_\pi^\omega(\varnothing)) \,\mathrm{d}G(\omega), \tag{S-IC}$$

$$\pi^* \in \operatorname*{argmax}_{\pi' \in \Pi} \int_\Theta \int_{M \cup \{\varnothing\}} \int_\Omega \alpha_{\pi'}^\omega(m) \,\mathrm{d}G(\omega) \,\mathrm{d}\ell_{\pi',\rho,\delta}(m|\theta) \,\mathrm{d}\overline{F}(\theta), \tag{Ex-Ante}$$

The definition of a covert-acquisition equilibrium, or *c-equilibrium*, is equivalent, except condition (o-Bayes) is replaced with

$$\beta_\pi \text{ is derived from Bayes rule given } \mu_0 \text{ and } \begin{cases} \ell_{\pi,\rho,\delta}, & \text{on } M, \\ \ell_{\pi^*,\rho,\delta}, & \text{on } \{\varnothing\}. \end{cases} \tag{c-Bayes}$$

That is, in contrast to the overt case, R's beliefs depend only on the on-path S's choice of π^* in the event of non-disclosure as she then cannot detect S's ex-ante deviations.

A.2 Proofs

A.2.1 Proof of Lemma 1

To obtain (2), integrate by parts to obtain

$$\mathbb{E}_F(x|x \leqslant x_\varnothing) = \frac{1}{F(x_\varnothing)} \int_0^{x_\varnothing} x \, \mathrm{d}F(x) = x_\varnothing - \frac{I(x_\varnothing)}{F(x_\varnothing)},$$

then plug it into (1) and rearrange the terms.

Next, we establish (i). By Lemma 1, the set of solutions is given by the roots of a function

$$\begin{aligned} \xi_{\rho,I} \colon [0,1] &\to \mathbb{R} \\ x &\mapsto \rho I(x) + (1-\rho)(x - x_0). \end{aligned}$$

If $\rho \neq 1$ or $\min \operatorname{supp} I = 0$, then it is continuous, strictly increasing, and has $\xi_{\rho,I}(0) \geqslant 0$ and $\xi_{\rho,I}(x_0) \leqslant 0$, with the first (second) inequality strict if $\rho = 1$ ($\min \operatorname{supp} I = 0$). Thus, it has a unique root by the Intermediate Value Theorem.

If $\rho = 1$, then $\xi_{\rho,I} = I$ equals zero on $[0, \min \operatorname{supp} I]$ and is strictly increasing on $[\min \operatorname{supp} I, 1]$. Therefore, $d_{1,I} = \min \operatorname{supp} I$ is unique root of I in $\operatorname{supp} I$ (which also establishes (iii)).

Finally, (ii) holds because $\xi_{\rho,I}(x)$ is strictly increasing in ρ and increasing in I (with respect to $\geqslant$) for all x.

A.2.2 Proof of Lemma 2

First, note that equilibrium conditions (R-IC) and (S-IC) can be viewed as maximizations of linear functions on $[0,1]$ and are, therefore, equivalent to

$$\alpha_\pi^\omega(m) = \mathbf{1}(\mathbf{E}\beta_\pi(m) \geqslant \omega) \tag{R-IC$'$}$$

for all $m \in M$ and $\omega \in \Omega$ (except, possibly, for $\mathbf{E}\beta_\pi(m) = \omega$), and

$$\delta_\pi(m) = \mathbf{1}(\mathbf{E}\beta_\pi(m) \geqslant \mathbf{E}\beta_\pi(\varnothing)) \tag{S-IC$'$}$$

for all $m \in M$ (except, possibly, when $\mathbf{E}\beta_\pi(m) = \mathbf{E}\beta_\pi(\varnothing)$).

Then, one can rewrite the (Ex-Ante) objective function as

$$\begin{aligned}
&\int_\Theta \int_{M\cup\{\varnothing\}} \int_\Omega \alpha^\omega_{\pi'}(m)\,\mathrm{d}G(\omega)\,\mathrm{d}\ell_{\pi,\rho,\delta}(m|\theta)\,\mathrm{d}\overline{F}(\theta)\\
&= \int_\Theta \int_{M\cup\{\varnothing\}} G(\mathbf{E}\beta_{\pi'}(m))\,\mathrm{d}\ell_{\pi',\rho,\delta}(m|\theta)\,\mathrm{d}\overline{F}(\theta)\\
&= \rho\int_\Theta\int_M G(\mathbf{E}\beta_{\pi'}(m))\,\mathrm{d}\left[\delta_{\pi'}(m)\pi'(m|\theta)\right]\,\mathrm{d}\overline{F}(\theta)\\
&\quad + G(\mathbf{E}\beta_{\pi'}(\varnothing))\left(1-\rho+\rho\int_\Theta\int_M(1-\delta_{\pi'}(m))\,\mathrm{d}\pi'(m|\theta)\,\mathrm{d}\overline{F}(\theta)\right)\\
&= \rho\int_\Theta\int_{\{m\in M:\ \mathbf{E}\beta_{\pi'}(m)>\mathbf{E}\beta_{\pi'}(\varnothing)\}} G(\mathbf{E}\beta_{\pi'}(m))\,\mathrm{d}\pi'(m|\theta)\,\mathrm{d}\overline{F}(\theta)\\
&\quad + G(\mathbf{E}\beta_{\pi'}(\varnothing))\left(1-\rho+\rho\int_\Theta \pi'(\{m\in M:\ \mathbf{E}\beta_{\pi'}(m)\leqslant\mathbf{E}\beta_{\pi'}(\varnothing)\}|\theta)\,\mathrm{d}\overline{F}(\theta)\right).
\end{aligned}$$

By the definition of $F_{\pi'}$, the (Ex-Ante) objective function can be further rewritten as

$$\rho\int_{\mathbf{E}\beta_{\pi'}(\varnothing)}^1 G(x)\,\mathrm{d}F_{\pi'} + G(\mathbf{E}\beta_{\pi'}(\varnothing))\left[1-\rho+\rho F_{\pi'}(\mathbf{E}\beta_{\pi'}(\varnothing))\right] = \nu_\rho(I_{\pi'}|\mathbf{E}\beta_{\pi'}(\varnothing))$$

To sum up, $(\pi^*,\delta,\alpha,\beta)$ satisfies (R-IC), (S-IC), and (Ex-Ante) if and only if it satisfies (R-IC′) and (S-IC′), and

$$\pi^* \in \operatorname*{argmax}_{\pi'\in\Pi} \nu_\rho(I_{\pi'}|\mathbf{E}\beta_{\pi'}(\varnothing)). \tag{Ex-Ante′}$$

Now consider the overt case. Since (o-Bayes) implies $\mathbf{E}\beta_{\pi'}(\varnothing) = d_{\rho,I_{\pi'}}$ for all $\pi'\in\Pi$, if I^* is an o-equilibrium structure then

$$I^* \in \operatorname*{argmax}_{I\in\mathcal{I}} \nu_\rho(I|d_{\rho,I}),$$

which is exactly the (Overt) program.

Next, consider the covert case. Note that (c-Bayes) implies, for all $\pi''\in\Pi$, $\mathbf{E}\beta_{\pi''}(\varnothing) = d_{\rho,I_{\pi^*}}$. Hence, if I^* is a c-equilibrium structure then

$$I^* \in \operatorname*{argmax}_{I\in\mathcal{I}} \nu_\rho(I|d_{\rho,I^*}),$$

which is exactly the (Covert) program.

Finally, to show the sufficiency of the two programs for the corresponding equilibria, we use $\{I_{\pi'}:\pi'\in\Pi\} = \mathcal{I}$. Given a solution I^* to the (Overt) (respectively, Covert) program, take any $\pi^*\in\Pi$ such that $I_{\pi^*} = I^*$, any β_π satisfying (o-Bayes) (respectively, c-Bayes) and let α and δ be defined as in (R-IC′) and (S-IC′) to construct a profile $(\pi^*,\delta,\alpha,\beta)$ which is an o-equilibrium (c-equilibrium, respectively) by the above arguments. □

A.2.3 Proof of Corollary 1

Endow $\mathcal{I}$ with a topology corresponding (induced under $\mu \mapsto (t \mapsto \int_0^t \mu[0,x]\mathrm{d}x)$) to the weak* topology on $\Delta[0,1]$. Then, the (Overt) program admits a solution since it has a compact domain and a continuous objective.

For the covert case, consider the correspondence

$$\begin{aligned}\Phi\colon \mathcal{I}\times[0,1] &\rightrightarrows \mathcal{I}\times[0,1]\\ (I,x_\varnothing) &\mapsto \operatorname*{argmax}_{I'\in\mathcal{I}} \nu_\rho(I'|x_\varnothing)\times\{d_{\rho,I}\}.\end{aligned}$$

Note that $x_\varnothing \mapsto \operatorname{argmax}_{I'\in\mathcal{I}} \nu_\rho(I'|x_\varnothing)$ is non-empty-, convex-, compact-valued and upper-hemicontinuous by Berge's Theorem and $I \mapsto d_{\rho,I}$ is a continuous mapping as follows from Lemma 1. Therefore, Φ is a Kakutani (non-empty-, convex-, compact-valued and upper-hemicontinuous) correspondence on a compact and convex domain and, therefore, it admits a fixed point by the Kakutani-Glicksberg-Fan theorem. □

A.2.4 Proof of Lemma 3

We first establish the following result which implies both the equivalence between pivoting and S-improvements of Lemma 3 and the equivalence between strict informativeness and (ex-ante) R-improvements.

Lemma 6. *Let $z_1,\dots,z_{2k-1}\in[0,1]$ and $Z_i^+ := [z_{2i}, z_{2i-1}], Z_i^- := [z_{2i-1}, z_{2i}]$ for all $i = 1,\dots,k$, where $z_0 := 0, z_{2k} = 1$. Then, for any $I, J \in \mathcal{I}$ the following statements are equivalent:*

(i) J is weakly above (below) I on each Z_i^+ (Z_i^-) and $J \neq I$,

(ii) $\int_0^1 (J-I)\,\mathrm{d}h > 0$ for all $h\colon[0,1]\to\mathbb{R}$ strictly increasing (decreasing) on each Z_i^+ (Z_i^-).

Proof. First, to show (i) implies (ii), suppose J is weakly above (below) I on each Z_i^+ (Z_i^-) and let $h\colon[0,1]\to\mathbb{R}$ be strictly increasing (decreasing) on each Z_i^+ (Z_i^-). Then, $\int_0^1 (J-I)\,\mathrm{d}h \geqslant \int_N (J-I)\,\mathrm{d}h$ for any interval $N\subseteq[0,1]$. If $J\neq I$, then there exists $\varepsilon>0, i\in\{0,\dots,2k\}$ and $0\leqslant \underline{z}<\overline{z}\leqslant 1$ such that either $J-I>\varepsilon$ on $[\underline{z},\overline{z}]\subseteq Z_i^+$ or $I-J>\varepsilon$ on $[\underline{z},\overline{z}]\subseteq Z_i^-$. In both cases, we have $h(\overline{z})\neq h(\underline{z})$ and hence

$$\int_0^1 (J-I)\,\mathrm{d}h \geqslant \int_{\underline{z}}^{\overline{z}} (J-I)\,\mathrm{d}h \geqslant \varepsilon|h(\overline{z})-h(\underline{z})| > 0.$$

Second, to show (ii) implies (i), suppose $\int_0^1 (J-I)\,\mathrm{d}h > 0$ for all $h\colon[0,1]\to\mathbb{R}$ strictly increasing (decreasing) on each Z_i^+ (Z_i^-), which immediately implies $J\neq I$. Next, take any

$x \in [0,1]$ and suppose $x \in Z_i^+$ for some i. Define $h^x := \mathbf{1}_{[x,1]} \colon [0,1] \to \mathbb{R}$ and consider a sequence of ramp functions $h_n^x := [1 - n(x - \mathbf{id})^+]^+ : [0,1] \to \mathbb{R}$. Note that $h_n^x \rightharpoonup h^x$ and, by assumption, $\int_0^1 (J - I)\, \mathrm{d}h^x > 0$ for all n. Hence,

$$J(x) - I(x) = \int_0^1 (J - I)\, \mathrm{d}h^x = \lim_{n\to\infty} \int_0^1 (J - I)\, \mathrm{d}h_n^x \geqslant 0$$

Finally, the case of $x \in Z_i^-$ for some i is analogous with $h^x := \mathbf{1}_{[0,x)}$ and $h_n^x := [1 - n(\mathbf{id} - x)^+]^+$.

□

Proof of Lemma 3. The result follows immediately from Lemma 6 by letting $k = 1, z_1 = \hat{\omega}$. □

Now note that Lemma 6 is related to the following strict version of the Blackwell-Rothschild-Stiglitz Theorem.

Corollary 6. *For all $I, J \in \mathcal{I}$, the following statements are equivalent*

(i) J is strictly more informative than I: $J > I$,

(ii) J is an R-improvement over I: $w(J) - w(I) = \int_0^1 (J - I)\, \mathrm{d}G > 0$ for all strictly increasing G.

Proof. The result follows immediately from Lemma 6 by letting $k = 1, z_1 = 1$. □

A.2.5 Proof of Lemma 4

Take any $I \in \mathcal{I}$. By Lemma 5 in Lipnowski et al. (2021), we can find some $t \in [0, \hat{\omega}]$, $\theta \in [\hat{\omega}, 1]$, and $J \in \mathcal{I}$ such that

- $J = \bar{I}$ on $[0, t]$, J is affine on $[t, \theta]$, and J is affine with unit slope on $[\theta, 1]$,
- J is a pivoted I or $J = I$.

Because the first part simply states that J is a t upper censorship, by Lemma 3, this implies that either J coicides with or is an S-improvement of I.

□

A.2.6 Proof of Corollary 2

First, note that by Lemma 4, any (FC) solution is an upper censorship since otherwise it could be S-improved. Therefore, a t upper censorship I_t solves (FC) if and only if t maximizes $\underline{v}_0 := t \mapsto v(I_t)$ over $[0,1]$. As follows from Lemma 7 (below) part (SQC) with $x_\varnothing = 0$, there is a unique maximizer $t_1^* \in (0, \hat{\omega})$.

Next, as $v_1(I|d_{1,I}) = v(I_1^D) = v(I)$, the t_1^* upper censorship is also the unique o-equilibrium structure.

Finally, by Lemma 1, $d_{1,I_1^*} = \min \operatorname{supp} I_1^* = 0$, and so for any $I \in \mathcal{I}$, $v_1(I|d_{1,I_1^*}) = v_1(I|0) = v(I)$. Hence, the t_1^* upper censorship is the unique c-equilibrium structure with 0 being the lowest point in the support. Now, by contradiction, suppose there is some (Covert) solution I with $x := \min \operatorname{supp} I > 0$. This implies I cannot be an upper censorship and so, by Lemma 4, there exists its S-improvement upper censorship J so that $v(J) > v(I)$. Also, $J|_{[0,x]} \geqslant 0 = I_{[0,x]}$. But then

$$v_1(J|d_{1,I}) - v_1(I|d_{1,I}) = v_1(J|x) - v_1(I|x) = v(J) - v(I) + \int_0^x J' \, \mathrm{d}G > 0,$$

which means J is a strictly profitable deviation.

□

A.2.7 Proof of Corollary 4

Suppose I^* is an o-equilibrium which satisfies (MP) and take any $I \in \mathcal{I}$. We will show that I is a o-equilibrium satisfying (MP) if and only if it is a c-equilibrium. By Lemma 5, it is weakly beneficial to covertly deviate from I to I^* because it is weakly beneficial to overtly deviate from I to I^* and there is an additional non-negative benefit since I^* satisfies the minimum principle. But then I is a c-equilibrium if and only if both of these non-negative benefits are zero which is equivalent to I being an o-equilibrium satisfying (MP).

□

A.2.8 Towards the Proofs of Theorems 1 and 2

In this section, we consider the relaxed maximization over upper censorship thresholds. We establish some properties of its objective function and introduce some notation which will be used in the proofs of the main results.

Properties of upper censorships. Fix any $t \in \Theta$. Let I_t denote the t upper censorship of $\bar{I}$, $F_t := I_t'$ be the corresponding CDF, $x_t := \int_t^1 \theta \, \mathrm{d}\bar{F}(\theta)/(1 - \bar{F}(t))$ be the conditional mean of

the upper pooling. That is,

$$I_t(x) = \begin{cases} \bar{I}(x), & x \leqslant t, \\ \bar{I}(t) + \overline{F}(t)(x-t), & x > t, \end{cases} \quad F_t(x) = \begin{cases} \overline{F}(x), & x \leqslant t, \\ \overline{F}(t), & x \in (t, x_t), \\ 1, & x \geqslant x_t, \end{cases} \quad x_t = \frac{x_0 + \bar{I}(t) - t\overline{F}(t)}{1 - \overline{F}(t)}$$

Given our assumptions on $\bar{I}$, x_t for $t = 0$ defined this way is consistent with our notation x_0 for the prior mean of $\bar{I}$. Moreover, $x_t > t$ for all $t \in [0,1)$ and x_t is strictly increasing in t. In particular, this implies that, by the Intermediate Value Theorem, there exists a unique $\underline{t} \in (0, \hat{\omega})$ such that

$$x_{\underline{t}} = \hat{\omega}.$$

Clearly, $I_t(x)$, $F_t(x)$, and x_t are almost everywhere continuous and differentiable in t. In particular,

$$\begin{aligned} \tfrac{\mathrm{d}}{\mathrm{d}t} I_t(x) &= \mathbf{1}_{[t,x_t]}(x)\bar{f}(t)(x-t), && \text{for all } (t,x) \in [0,1]^2 \text{ such that } x \neq x_t, \\ \tfrac{\mathrm{d}}{\mathrm{d}t} F_t(x) &= \mathbf{1}_{[t,x_t]}(x)\bar{f}(t), && \text{for all } (t,x) \in [0,1]^2 \text{ such that } x \notin \{t, x_t\}, \\ \tfrac{\mathrm{d}}{\mathrm{d}t} x_t &= \frac{(x_t - t)\bar{f}(t)}{1 - \overline{F}(t)}, && \text{for all } t \in [0,1). \end{aligned}$$

Next, fix any $\rho \in (0,1]$. Then, the Bayes-consistent non-disclosure posterior equals

$$d_{\rho,I_t} = \begin{cases} \dfrac{\rho[\overline{F}(t)t - \bar{I}(t)] + (1-\rho)x_0}{1 - \rho(1 - \overline{F}(t))}, & \text{if } t \leqslant d_{\rho,\bar{I}}, \\ d_{\rho,\bar{I}}, & \text{if } t > d_{\rho,\bar{I}}. \end{cases}$$

and is differentiable in t and ρ everywhere except at $(t,\rho) = (0,1)$ with

$$\begin{aligned} \tfrac{\mathrm{d}}{\mathrm{d}t} d_{\rho,I_t} &= -\frac{\rho\bar{f}(t)\,[d_{\rho,I_t} - t]^+}{1 - \rho(1 - \overline{F}(t))} \leqslant 0, \\ \tfrac{\mathrm{d}}{\mathrm{d}\rho} d_{\rho,I_t} &= -\frac{I_t(d_{\rho,I_t}) + x_0 - d_{\rho,I_t}}{1 - \rho(1 - \overline{F}(t) \wedge \overline{F}(d_{\rho,I_t}))} < 0, \end{aligned}$$

and satisfies

$$\begin{aligned} t > d_{\rho,\bar{I}} &\iff t > d_{\rho,I_t}, \\ t < d_{\rho,\bar{I}} &\iff t < d_{\rho,I_t}. \end{aligned}$$

For convenience, we extend this mapping by continuity: $d_{0,I_t} \coloneqq \lim_{\rho \to 0} d_{0,I_t} = x_0$.

Properties of the relaxed (FC) objective with the truncated G. We begin with the following lemma which will be key in both overt and covert cases. For any $t \in \Theta$, let $\underline{v}_{x_\varnothing}(t)$ denote the expected full-commitment S payoff from choosing ICDF I_t of R's posterior means given optimal disclosure and some fixed R's non-disclosure posterior $x_\varnothing \in [0, x_0]$, that is define

$$\begin{aligned}&\underline{v}\colon \Theta \times [0,x_0] \to \mathbb{R},\\ &\qquad (t, x_\varnothing) \mapsto \underline{v}_{x_\varnothing}(t) \coloneqq \int_0^1 G_{\vee x_\varnothing}\,\mathrm{d}F_t - G(x_0)\end{aligned}$$

where $G_{\vee x_\varnothing}(x) \coloneqq \max\{G(x), G(x_\varnothing)\}$ for all $x \in X$. Note that $\underline{v}_{x_\varnothing}$ is exactly what S is maximizing when she is best-responding to R's non-disclosure belief in the covert case. In addition, for the overt case, we will use the fact that the function $\underline{v}_0(t) = v(I_t)$ coincides with the (relaxed) full-commitment objective.

Lemma 7. *The function $\underline{v}$ has the following properties:*

(Cont) $\underline{v}_{x_\varnothing}$ *is continuous for all* $x_\varnothing \in [0, x_0]$,

(Incr) $\underline{v}_{x_\varnothing}$ *is strictly increasing on* $[0, \underline{t}]$ *for all* $x_\varnothing \in [0, x_0]$,

(SQC) $\underline{v}_{x_\varnothing}$ *is strictly quasiconcave with the peak in* $(0, \hat{\omega})$ *for all* $x_\varnothing \in [0, x_0]$,

(ZMD) $\underline{v}$ *has zero marginal differences on* $\{(t, x_\varnothing) \in [0,1] \times [0, x_0] \colon x_\varnothing < t\}$,

(SIMD) $\underline{v}$ *has strictly increasing marginal differences on* $\{(t, x_\varnothing) \in (0,1) \times (0, x_0) \colon x_\varnothing > t\}$.

Proof. First, (Cont) holds since $\overline{F}, G, t \mapsto x_t$ are continuous and we have

$$\underline{v}_{x_\varnothing}(t) = \int_0^t G_{\vee x_\varnothing} \bar{f} + (1 - \overline{F}(t)) G(x_t) - G(x_0).$$

Second, (ZMD) and (SIMD) both follow from the observation that

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}t\,\mathrm{d}x_\varnothing}\underline{v}_{x_\varnothing}(t) &= \frac{\mathrm{d}^2}{\mathrm{d}t\,\mathrm{d}x_\varnothing}\int_0^1 G_{\vee x_\varnothing}\,\mathrm{d}F_t\\ &= \frac{\mathrm{d}^2}{\mathrm{d}t\,\mathrm{d}x_\varnothing}\left[F_t(x_\varnothing)G(x_\varnothing) + \int_{x_\varnothing}^1 G\,\mathrm{d}F_t\right]\\ &= \frac{\mathrm{d}^2}{\mathrm{d}t\,\mathrm{d}x_\varnothing}\left[1 - \int_{x_\varnothing}^1 F_t\,\mathrm{d}G\right]\\ &= \frac{\mathrm{d}}{\mathrm{d}t}\left[F_t(x_\varnothing)g(x_\varnothing)\right]\\ &= \mathbf{1}_{[t,x_t]}(x_\varnothing)\bar{f}(t)g(x_\varnothing).\end{aligned}$$

which is strictly positive if $0 < t < x_\varnothing < x_0$ and zero if $x_\varnothing < t$.

Third, we show (Incr) and (SQC). Denote the tangent lines to G and $G_{\vee x_\varnothing}$ at t as

$$G^T(s|t) \coloneqq G(t) + g(t)(s-t),$$
$$G^T_{\vee x_\varnothing}(s|t) \coloneqq G_{\vee x_\varnothing}(t) + \mathbf{1}_{[x_\varnothing,1]}(t)g(t)(s-t).$$

for all $s \in [0,1]$. Fix any $x_\varnothing \in [0, x_0]$ and note that strict convexity (concavity) of G below (above) $\hat{\omega}$ implies

$$\text{for all } t,s \in [0,\hat{\omega}], t \neq s: \quad G(s) > G^T(s|t),$$
$$\text{for all } t,s \in [\hat{\omega},1], t \neq s: \quad G(s) < G^T(s|t),$$

and, therefore,

$$\text{for all } t,s \in [0,\hat{\omega}], s \neq t \qquad G_{\vee x_\varnothing}(s) \geqslant G^T_{\vee x_\varnothing}(s|t), \qquad \text{(wConvexity)}$$
$$\text{for all } t,s \in [0,\hat{\omega}], s \neq t, s \vee t \geqslant x_\varnothing \qquad G_{\vee x_\varnothing}(s) > G^T_{\vee x_\varnothing}(s|t), \qquad \text{(sConvexity)}$$
$$\text{for all } t,s \in [\hat{\omega},1], s \neq t: \qquad G_{\vee x_\varnothing}(s) < G^T_{\vee x_\varnothing}(s|t). \qquad \text{(sConcavity)}$$

By definition of $\underline{v}_{x_\varnothing}$, we have for all $t \neq x_\varnothing$

$$\underline{v}_{x_\varnothing}(t) = \int_0^1 G_{\vee x_\varnothing}\,\mathrm{d}F_t - G(x_0) = \int_0^t G_{\vee x_\varnothing}\,\mathrm{d}\overline{F} + (1-\overline{F}(t))G(x_t) - G(x_0),$$
$$\underline{v}'_{x_\varnothing}(t) = \bar{f}(t)\big[G_{\vee x_\varnothing}(t) - G(x_t) - g(x_t)(t-x_t)\big] = \bar{f}(t)\big[G_{\vee x_\varnothing}(t) - G^T(t|x_t)\big].$$

Thus, to show (Incr), it is sufficient to show that $G_{\vee x_\varnothing}(t) > G^T(t|x_t)$ for all $t \in [0,\underline{t}]$. But $t \in [0,\underline{t}]$ implies $0 \leqslant t < x_t \leqslant \hat{\omega}$, and, hence, the desired inequality holds for all $t \in [0,\underline{t}]$ by (sConvexity).

Similarly, to prove (SQC), it is sufficient to show that for all $0 < t_1 < t_2 < 1$,

$$G_{\vee x_\varnothing}(t_2) \geqslant G^T_{\vee x_\varnothing}(t_2|y_2) \implies G_{\vee x_\varnothing}(t_1) > G^T_{\vee x_\varnothing}(t_1|y_1), \tag{3}$$
$$G_{\vee x_\varnothing}(t_1) \leqslant G^T_{\vee x_\varnothing}(t_1|y_1) \implies G_{\vee x_\varnothing}(t_2) < G^T_{\vee x_\varnothing}(t_2|y_2), \tag{4}$$

where $y_1 \coloneqq x_{t_1} < x_{t_2} =: y_2$.

To prove (3) and (4), consider the following exhaustive cases:

1. Suppose $y_1 \leqslant \hat{\omega}$. Then, $t_1, y_1 \in (0,\hat{\omega}]$ and $t_1 \neq y_1 > x_0 \geqslant x_\varnothing$. Hence, both (3) and (4) hold since (sConvexity) implies the conlusion of (3) always holds and the premise of (4) never holds.

2. Suppose $t_2 \geqslant \hat{\omega}$. Then, $t_2, y_2 \in [\hat{\omega}, 1)$ and $y_2 > t_2$. Hence, both (3) and (4) hold since (sConcavity) implies the conlusion of (4) always holds and the premise of (3) never holds.
3. Suppose $t_2 < \hat{\omega} < y_1$. Then, $0 < t_1 < t_2 < \hat{\omega}$ and $x_\varnothing \leqslant x_0 \leqslant \hat{\omega} < y_1 < y_2 \leqslant 1$. First, to prove (3), suppose $G_{\vee x_\varnothing}(t_2) \geqslant G^T_{\vee x_\varnothing}(t_2|y_2)$. Note that

$$g(y_1) > g(y_2) > \frac{G(\hat{\omega}) - G_{\vee x_\varnothing}(t_2)}{\hat{\omega} - t_2} > g(t_2), \tag{5}$$

where the first inequality follows monotonicity of g on $[\hat{\omega}, 1]$, the third — from (sConvexity) for $t = t_2, s = \hat{\omega}$ and the second — from the premise of (3) and (sConcavity) for $t = y_2, s = \hat{\omega}$ as $G_{\vee x_\varnothing}(t_2) - g(y_2)(t_2 - y_2) \geqslant G(y_2) > G(\hat{\omega}) - g(y_2)(\hat{\omega} - y_2)$. Now, the conclusion of (3) follows from

$$\begin{aligned}
G_{\vee x_\varnothing}(t_1) &\geqslant g(t_2)(t_1 - t_2) + G_{\vee x_\varnothing}(t_2) && \text{by (wConvexity) for } t = t_2, s = t_1 \\
&\geqslant g(t_2)(t_1 - t_2) + G(y_2) + g(y_2)(t_2 - y_2) && \text{by the premise of (3)} \\
&> g(y_2)(t_1 - t_2) + G(y_2) + g(y_2)(t_2 - y_2) && \text{by (5)} \\
&= g(y_2)(t_1 - y_2) + G(y_2) \\
&> g(y_2)(t_1 - y_2) + G(y_1) - g(y_2)(y_1 - y_2) && \text{by (sConcavity) for } t = y_2, s = y_1 \\
&= g(y_2)(t_1 - y_1) + G(y_1) \\
&> g(y_1)(t_1 - y_1) + G(y_1). && \text{by (5)}
\end{aligned}$$

Finally, to prove (4), suppose $G_{\vee x_\varnothing}(t_1) \leqslant G^T_{\vee x_\varnothing}(t_1|y_1)$ and note that

$$\frac{G(\hat{\omega}) - G_{\vee x_\varnothing}(t_1)}{\hat{\omega} - t_1} > g(y_1) > g(y_2), \tag{6}$$

where the second inequality follows from monotonicity of g on $[\hat{\omega}, 1]$ and the first — from the premise of (4) and (sConcavity) for $t = y_1, s = \hat{\omega}$ as $G_{\vee x_\varnothing}(t_1) - g(y_1)(t_1 - y_1) \geqslant$

$G(y_1) > G(\hat{\omega}) - g(y_1)(\hat{\omega} - y_1)$. Thus, the conclusion of (4) follows from

$$
\begin{aligned}
& g(y_2)(t_2 - y_2) + G(y_2) \\
&> g(y_2)(t_2 - y_2) + G(y_1) - g(y_2)(y_1 - y_2) && \text{by (sConcavity) for } t = y_2, s = y_1 \\
&= g(y_2)(t_2 - y_1) + G(y_1) \\
&> g(y_1)(t_2 - y_1) + G(y_1) && \text{by (6)} \\
&> g(y_1)(t_2 - y_1) + G(\hat{\omega}) - g(y_1)(\hat{\omega} - y_1) && \text{by (sConcavity) for } t = y_1, s = \hat{\omega} \\
&= g(y_1)(t_2 - \hat{\omega}) + G(\hat{\omega}) \\
&> \frac{G(\hat{\omega}) - G_{\vee x_\varnothing}(t_1)}{\hat{\omega} - t_1}(t_2 - \hat{\omega}) + G(\hat{\omega}) && \text{by (6)} \\
&> G_{\vee x_\varnothing}(t_2). && \text{chordal slopes} \uparrow\text{: } \tfrac{G(\hat{\omega}) - G_{\vee x_\varnothing}(t_2)}{\hat{\omega} - t_2} > \tfrac{G(\hat{\omega}) - G_{\vee x_\varnothing}(t_1)}{\hat{\omega} - t_1}
\end{aligned}
$$

Finally, the peak is positive by (Incr) and strictly below $\hat{\omega}$ since $\underline{\nu}'_{x_\varnothing}(s) = G_{\vee x_\varnothing}(t) < G(x_t) - g(x_t)(x_t - s)$ for all $s \geqslant \hat{\omega}$ by (sConcavity). □

Properties of the relaxed (Overt) objective. Now consider the optimization problem

$$\max_{t \in \Theta} \tilde{\nu}_\rho(t), \tag{Overt′}$$

where

$$
\begin{aligned}
\tilde{\nu} \colon \Theta \times (0,1] &\to \mathbb{R}, \\
(t, \rho) &\mapsto \tilde{\nu}_\rho(t) := \frac{\nu_\rho(I_t | d_{\rho, I_t})}{\rho}.
\end{aligned}
$$

Lemma 8. *The function $\tilde{\nu}$ has the following properties:*

(CD) $\tilde{\nu}_\rho$ *is continuous and a.e. differentiable for all* $\rho \in (0,1]$,

(Incr) $\tilde{\nu}_\rho$ *is strictly increasing on* $[0, \underline{t}]$ *for some* $\underline{t} \in (0, \hat{\omega})$ *and all* $\rho \in (0,1]$,

(SQC) $\tilde{\nu}_1$ *is strictly quasiconcave with the peak in* $(0, \hat{\omega})$,

(Diff) $\tilde{\nu}_1 - \tilde{\nu}_\rho$ *is strictly increasing on* $[0, d_{\rho, \bar{I}}]$, *and constant on* $[d_{\rho, \bar{I}}, 1]$.

(ID) $\tilde{\nu}$ *has increasing differences,*

(SIMD) $\tilde{\nu}$ *has strictly increasing marginal differences on* $(0, d_{\rho', \bar{I}}) \times (0, \rho']$ *for all* $\rho' \in (0,1]$,

Proof. First, (CD) follows from continuity and a.e. differentiability of $I_t(x)$ and d_{ρ, I_t} in t.

Second, (Incr) and (SQC) follows directly from (Incr) and (SQC) of Lemma 7 and the fact that $\underline{\nu}_0 = \tilde{\nu}_1$.

Third, we establish (Diff). We have

$$\begin{aligned}\tilde{v}_\rho(t) &= \frac{v_\rho(I_t|d_{\rho,I_t})}{\rho}\\ &= \frac{v\left([I_t]^D_\rho\right)}{\rho}\\ &= \frac{1}{\rho}\int_0^1 \Big(\big[\rho I_t(x) + (1-\rho)(x-x_0)\big]^+ - \underline{I}(x)\Big)\,\mathrm{d}g(x)\\ &= \int_{d_{\rho,I_t}}^1 \left[I_t(x) + \frac{(1-\rho)}{\rho}(x-x_0) - \frac{\underline{I}(x)}{\rho}\right]\mathrm{d}g(x).\end{aligned}$$

and hence by Lemma 1

$$\begin{aligned}\tilde{v}_1(t) - \tilde{v}_\rho(t) &= \int_0^1 (I_t - \underline{I})\,\mathrm{d}g - \int_{d_{\rho,I_t}}^1 \left[I_t(x) + \frac{(1-\rho)}{\rho}(x-x_0) - \frac{\underline{I}(x)}{\rho}\right]\mathrm{d}g(x)\\ &= \int_0^{d_{\rho,I_t}} I_t\,\mathrm{d}g + \int_{d_{\rho,I_t}}^{x_0} \frac{(1-\rho)}{\rho}(x_0-x)\,\mathrm{d}g(x)\\ &= \int_0^{x_0} I_t(x) \vee \frac{(1-\rho)}{\rho}(x_0-x)\,\mathrm{d}g(x).\end{aligned}$$

Thus, $\tilde{v}_1 - \tilde{v}_\rho$ is strictly increasing on $[0, d_{\rho,\bar{I}}]$ and constant on $[d_{\rho,\bar{I}}, 1]$ because so is $t \mapsto I_t(x) \vee \frac{(1-\rho)}{\rho}(x_0 - x) = I_t(x)$ for all $x \in [t, d_{\rho,I_t}]$ (since $d_{\rho,I_t} > d_{\rho,\bar{I}} > t$ if $t < d_{\rho,\bar{I}}$ and $d_{\rho,I_t} = d_{\rho,\bar{I}}$ otherwise).

Finally, notice that (ID) and (SIMD) are equivalent to $\tilde{v}_1 - \tilde{v}_\rho$ having increasing differences everywhere and strictly increasing marginal differences on $(0, d_{\tilde{\rho},\bar{I}}) \times (0, \tilde{\rho}]$ for all $\tilde{\rho} \in (0,1]$. To prove this, consider

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}\rho\,\mathrm{d}t}\big[\tilde{v}_\rho(t) - \tilde{v}_1(t)\big] &= -\frac{\mathrm{d}^2}{\mathrm{d}\rho\,\mathrm{d}t}\left[\int_0^{d_{\rho,I_t}} I_t\,\mathrm{d}g + \int_{d_{\rho,I_t}}^{x_0} \frac{(1-\rho)}{\rho}(x_0-x)\,\mathrm{d}g(x)\right]\\ &= -\frac{\mathrm{d}}{\mathrm{d}\rho}\left[\int_t^{d_{\rho,I_t}} \bar{f}(t)(x-t)\,\mathrm{d}g(x) + \frac{\mathrm{d}d_{\rho,I_t}}{\mathrm{d}t}\left(I_t(d_{\rho,I_t}) - \frac{(1-\rho)}{\rho}(x_0 - d_{\rho,I_t})\right)\right]\\ &= -\frac{\mathrm{d}d_{\rho,I_t}}{\mathrm{d}\rho}\bar{f}(t)(d_{\rho,I_t} - t)^+ g'(d_{\rho,I_t}),\end{aligned}$$

where $I_t(d_{\rho,I_t}) - \frac{(1-\rho)}{\rho}(x_0 - d_{\rho,I_t}) = 0$ by the definition of d_{ρ,I_t}. Therefore, $\tilde{v}_\rho$ has the desired properties because, for all $\tilde{\rho} \in (0,1]$, $t \in (0,1)$, the terms $g'(d_{\tilde{\rho},I_t})$, $-\frac{\mathrm{d}d_{\tilde{\rho},I_t}}{\mathrm{d}\tilde{\rho}}$ and $d_{\tilde{\rho},I_t} - t$ are (strictly) positive (for $d_{\tilde{\rho},\bar{I}} > t$). □

A.2.9 Proof of Theorem 1

Denote the solution correspondence of the (Overt′) program as

$$T^o\colon (0,1] \to [0,1]$$
$$\rho \mapsto \operatorname*{argmax}_{t\in[0,1]} \tilde{\nu}_\rho(t),$$

and note that $T^o_\rho := T^o(\rho) = \operatorname{argmax}_{[0,1]} \tilde{\nu}_\rho = \operatorname{argmax}_{[0,1]} \rho\tilde{\nu}_\rho = \operatorname{argmax}_{t\in[0,1]} \nu_\rho(I_t|d_{\rho,I_t})$. Hence, by Lemma 2 and Corollary 3, I is an o-equilibrium evidence structure if and only if I is disclosure equivalent to $I_{t^o_\rho}$ for some $t^o_\rho \in T^o_\rho$.

It is then sufficient to show that there exists $\overline{\rho}^o \in [0,1]$ such that

(i) T^o is non-empty-valued, compact-valued, and upper hemicontinuous,

(ii) T^o is increasing in the strong set order,

(iii) $T^o_\rho = \{t^*_1\}$ for all $\rho > \overline{\rho}^o$, where $\{t^*_1\} := \operatorname{argmax}_{[0,1]} \tilde{\nu}_1$,

(iv) $0 < t^o_\rho < d_{\rho,\bar{I}}$ for all $t^o_\rho \in T^o_\rho \setminus \{t^*_1\}$,

(v) $\overline{\rho}^o < 1$,

(vi) every selection from T^o is strictly increasing on $(0,\overline{\rho}^o]$.

Now we show that all these properties follow from the properties of $\tilde{\nu}$ shown in Lemma 8. First, (CD) implies (i) by Berge's Maximum Theorem. Second, (ID) implies (ii) by the Weak Monotone Comparative Statics Theorem (Topkis, 1978, Theorem 6.1). Third, since (SQC) implies $T^o_1 = \{t^*_1\}$, we define $\overline{\rho}^o := \inf\{\rho \in (0,1]\colon T^o_\rho = \{t^*_1\}\} \in [0,1]$ so that (iii) automatically holds.

Fourth, (iv) holds because (Incr) implies $\tilde{\nu}_\rho$ is strictly increasing below $\underline{t} > 0$ and (SQC) and (Diff) imply $\operatorname{argmax}_{[d_{\overline{\rho}^o,\bar{I}},1]} \tilde{\nu}_\rho = \{t^*_1\}$. Moreover, the same properties imply

$$T^o_\rho \cap [d_{\rho,\bar{I}},1] = \operatorname*{argmax}_{[0,1]} \tilde{\nu}_\rho \cap [d_{\rho,\bar{I}},1] \subseteq \operatorname*{argmax}_{[d_{\rho,\bar{I}},1]} \tilde{\nu}_\rho = \operatorname*{argmax}_{[d_{\rho,\bar{I}},1]} \tilde{\nu}_1 = \{t^*_1\}$$

and so for ρ close enough to 1, we have $t^*_1 \in [d_{\rho,\bar{I}},1]$ (since $d_{\rho,\bar{I}} \underset{\rho\to 1}{\to} d_{1,\bar{I}} = \min\operatorname{supp}\bar{I} = 0$) which implies (v) by upper hemicontinuity of T^o.

Finally, we prove (vi) by contradiction.[26] Suppose some selection from T^o is not strictly increasing on $(0,\overline{\rho}^o]$. Since T^o is increasing in the strong set order, this means there exist

[26] Although the logic here is very similar to Edlin and Shannon (1998), their Strict Monotonicity Theorem 1 is not directly applicable here due to the fact that the strictly increasing marginal differences property (SIMD) holds on a contracting domain $[0, d_{\rho,\bar{I}}]$.

$0 < \rho_1 < \rho_2 \leqslant \overline{\rho}^o, t \in T^o_{\rho_1} \cap T^o_{\rho_2}$. Since $t \in (0,1)$, we have $\tilde{\nu}'_{\rho_1}(t) = \tilde{\nu}'_{\rho_2}(t) = 0$. If $t \neq t^*_1$, then $t < d_{\rho,\bar{I}}$ by (iv) and so we get a contradiction with the implication of (SIMD)

$$0 = \tilde{\nu}'_{\rho_2}(t) - \tilde{\nu}'_{\rho_1}(t) = \int_{\rho_1}^{\rho_2} \frac{\mathrm{d}^2}{\mathrm{d}\rho\,\mathrm{d}t} \tilde{\nu}_\rho(t) \mathrm{d}\rho > 0.$$

If $t = t^*_1$, then by definition of $\overline{\rho}^o$, there exists $s \in T^o_{\rho_1} \setminus \{t\}$ such that $s < d_{\rho_2,\bar{I}} \leqslant t$ and so by (SIMD) we get

$$\tilde{\nu}_{\rho_1}(s) - \tilde{\nu}_{\rho_1}(t) \geqslant \tilde{\nu}_{\rho_2}(s) - \tilde{\nu}_{\rho_2}(t) + \int_{\rho_1}^{\rho_2} \int_s^t \frac{\mathrm{d}^2}{\mathrm{d}\rho\,\mathrm{d}t} \tilde{\nu}_\rho(t)\,\mathrm{d}t\,\mathrm{d}\rho > 0,$$

which is a contradiction with $t \in T^o_{\rho_1}$. □

A.2.10 Proof of Theorem 2

By Lemma 2 and Corollary 5, I is a c-equilibrium evidence structure if and only if I is disclosure equivalent to some I_{t^*} such that

$$t^* \in \operatorname*{argmax}_{t\in[0,1]} \nu_\rho(I_t | d_{\rho,I_{t^*}}). \tag{Covert′}$$

The proof of the theorem follows directly from the following two claims establishing the properties of the (Covert′) fixed-point program.

Claim 1. The S best response correspondence

$$\begin{aligned} BR\colon [0,x_0] &\to [0,1] \\ x_\varnothing &\mapsto \operatorname*{argmax}_{t\in[0,1]} \underline{\nu}_{x_\varnothing}(t). \end{aligned}$$

has the following properties:

(i) BR is a singleton-valued and, thus, can be treated as a function,

(ii) BR is continuous,

(iii) $BR(x_\varnothing) = t^*_1$ for all $x_\varnothing \in [0, t^*_1 \wedge x_0]$,

(iv) $BR(x_\varnothing) \in [t^*_1, t^*_1 \vee x_\varnothing]$ for all $x_\varnothing \in [t^*_1, x_0]$,

(v) BR is strictly increasing on $[t^*_1, x_0]$ if $t^*_1 < x_0$.

Proof. First, by Berge's Maximum Theorem, we have $|BR(x_\varnothing)| \geqslant 1$ and upper hemicontinuity of BR. Second, we have $|BR(x_\varnothing)| \leqslant 1$ due to strict quasiconcavity of $\underline{\nu}_{x_\varnothing}$. Thus, we have (i) and (ii).

Third, since by (ZMD) and (SIMD), we have $\underline{v}_{x_\varnothing} - \underline{v}_0$ strictly increasing on $[0, x_\varnothing]$ and constant on $[x_\varnothing, 1]$. Recall that $\underline{v}_0 = \tilde{v}_1$ is strictly quasiconcave with the peak $t_1^* \in (0, \hat{\omega})$. Therefore, $\underline{v}_{x_\varnothing} = (\underline{v}_{x_\varnothing} - \underline{v}_0) + \tilde{v}_1$ is strictly increasing on $[0, t_1^*]$ and strictly decreasing on $[t_1^* \vee x_\varnothing, 1]$, which immediately implies (iii) and (iv). In addition, this implies that maximizers are always interior, that is,

$$BR(x_\varnothing) \in [t_1^*, t_1^* \vee x_\varnothing] \subset (0, \hat{\omega} \vee x_0) \subseteq (0,1) \implies \underline{v}'_{x_\varnothing}(BR(x_\varnothing)) = 0 \text{ for all } x_\varnothing \in [0, x_0].$$

Fourth, to show (v), suppose, by contradiction, there exist $t_1^* \leqslant x_1 < x_2 \leqslant x_0$ such that $BR(x_1) \geqslant BR(x_2)$. Note that BR is weakly increasing by the Weak Monotone Comparative Statics Theorem (Topkis, 1978) because (ZMD) and (SIMD) imply increasing differences. Thus, $BR(x_1) = BR(x_2) = t \in (t_1^*, x_1)$ and $\underline{v}'_{x_1}(t) = \underline{v}'_{x_2}(t) = 0$. But then $\int_{x_1}^{x_2} \frac{\mathrm{d}^2}{\mathrm{d}x_\varnothing \, \mathrm{d}t} \underline{v}_{x_\varnothing}(t) = \underline{v}'_{x_2}(t) - \underline{v}'_{x_1}(t) = 0$ which contradicts (SIMD). □

Claim 2. There exists $\overline{\rho}^c \in [0,1]$ such that the fixed-point correspondence

$$\begin{aligned} T^c \colon (0,1] &\to [0,1] \\ \rho &\mapsto T^c_\rho := \{t \in [0,1] \colon t \in BR(d_{\rho, I_t})\} \end{aligned}$$

has the following properties:

(a) T^c is a singleton-valued and, thus, can be treated as a function,

(b) T^c is continuous,

(c) $T^c_\rho = t_1^*$ for all $\rho \in [\overline{\rho}^c, 1]$,

(d) $T^c_\rho < d_{\rho, \bar{I}}$ for all $\rho \in [\overline{\rho}^c, 1]$,

(e) T^c is strictly increasing on $(0, \overline{\rho}^c)$,

(f) $\overline{\rho}^c \in [0, \overline{\rho}^o]$.

Proof. Define $\overline{\rho}^c := \inf\{\rho \in (0,1] \colon d_{\rho, \bar{I}} \leqslant t_1^*\} \leqslant \overline{\rho}^o$. For any $\rho \in (0,1]$, define a function

$$\begin{aligned} \tilde{d}_\rho \colon [0,1] &\to [0, x_0) \\ t &\mapsto d_{\rho, I_t}. \end{aligned}$$

First, fix any $\rho \geqslant \overline{\rho}^c$ so that $d_{\rho, \bar{I}} \leqslant t_1^*$. Then, Claim 1 implies $T^c_\rho = \{t_1^*\}$ because

$$\begin{aligned} \{BR(d_{\rho, I_t}) : t \in [0, t_1^*]\} &= BR\left([d_{\rho, \bar{I}}, x_0]\right) \subseteq [t_1^*, 1], \\ \{BR(d_{\rho, I_t}) : t \in (t_1^*, 1]\} &= BR(t_1^* \wedge x_0) = \{t_1^*\}. \end{aligned}$$

Second, fix any $0 < \rho < \overline{\rho}^c$ so that $x_0 > d_{\rho,I_t} \geqslant d_{\rho,\bar{I}} > t_1^*$ for all $t \in [0,1]$. Then, we have $T_\rho^c \subseteq [t_1^*, x_0]$ since

$$\{BR(d_{\rho,I_t}) : t \in [0,1]\} = BR\left([d_{\rho,\bar{I}}, x_0]\right) \subseteq BR([t_1^*, x_0]) \subseteq [t_1^*, x_0].$$

Now let $\bar{t} = BR(x_0) > t_1^*$ and note T_ρ^c is the set of the roots of the function

$$\begin{aligned} \Delta_\rho \colon [t_1^*, \bar{t}] &\to \mathbb{R} \\ t &\mapsto BR^{-1}(t) - d_{\rho,I_t} \end{aligned}$$

is continuous, strictly increasing, and has $\Delta_\rho(t_1^*) = t_1^* - d_{\rho,I_{t_1^*}} < 0$ and $\Delta_\rho(\bar{t}) = x_0 - d_{\rho,\bar{I}} > 0$. Hence, by the Intermediate Value Theorem, there exists a unique root of Δ_ρ continuous in ρ. Moreover, since $\frac{\mathrm{d}}{\mathrm{d}\rho}\Delta_\rho = -\frac{\mathrm{d}}{\mathrm{d}\rho} d_{\rho,I_t} > 0$ for all $t \in [0,1], \rho \in (0,1]$, the root is strictly decreasing in ρ, which completes the proof of the claim. □

□

A.2.11 Proof of Proposition 1

The result follows from two observations. First, for any reliability level $\rho \in (0, \overline{\rho}^c)$ (which implies $\rho < \overline{\rho}^c \leqslant \overline{\rho}^o$ by Theorem 2), the c-equilibrium pass/fail threshold is strictly above the o-equilibrium pass/fail threshold:

$$\begin{aligned} t_\rho^c &> t_{\overline{\rho}^c}^c && \text{(Theorem 2: c-equilibrium threshold is decreasing in } \rho) \\ &= d_{\overline{\rho}^c,\bar{I}} && \text{(Theorem 2: by the definition of } \overline{\rho}^c) \\ &\geqslant d_{\overline{\rho}^o,\bar{I}} && \text{(Lemma 1: more disclosure under } \overline{\rho}^o \text{ than under } \overline{\rho}^c \leqslant \overline{\rho}^o) \\ &> t_\rho^o. && \text{(Theorem 1)} \end{aligned}$$

Second, we show that the information structure $[I_t]_\rho^D$ corresponding to the (Disclosed) pass/fail test I_t with a threshold t is strictly Blackwell-increasing in t for $t \in [0, d_{\rho,\bar{I}}]$. To see this, first note that one can write the ICDF of the t-threshold pass/fail test as

$$I_t(x) = \max\{0, \bar{I}(t) + \overline{F}(t)(x-t), x - x_0\}$$

and the corresponding (Disclosed) ICDF as

$$\begin{aligned} [I_t]_\rho^D(x) &= [\rho I_t(x) + (1-\rho)(x-x_0)]^+ \\ &= \max\{0, \rho(\bar{I}(t) + \overline{F}(t)(x-t)) + (1-\rho)(x-x_0), x - x_0\} \end{aligned}$$

Now take any $x \geqslant d_{\rho,\bar{I}} \geqslant t_2 > t_1$ and note that

$$\bar{I}(t_2) + \overline{F}(t_2)(x - t_2) \geqslant \bar{I}(t_2) + \overline{F}(t_1)(x - t_2) > \bar{I}(t_1) + \overline{F}(t_1)(x - t_1),$$

because $\overline{F}$ is strictly increasing and $\bar{I}$ is strictly convex. Therefore, $\rho(\bar{I}(t)+\overline{F}(t)(x-t))+(1-\rho)(x-x_0)$ is strictly increasing in t for $x \geqslant d_{\rho,\bar{I}} \geqslant t$ and, hence, $\max\{0, \rho(\bar{I}(t)+\overline{F}(t)(x-t))+(1-\rho)(x-x_0)\}$ is weakly (strictly) increasing for $t \leqslant d_{\rho,\bar{I}}$ and any $x \in [0,1]$ (any $x > d_{\rho,\bar{I}}$). This implies $[I_{t_2}]^D_\rho > [I_{t_1}]^D_\rho$ as desired.

References

Acharya, Viral V, Peter DeMarzo, and Ilan Kremer (2011): "Endogenous Information Flows and the Clustering of Announcements," *American Economic Review*, 101 (7), 2955–2979. 14

Ali, S Nageeb, Nima Haghpanah, Xiao Lin, and Ron Siegel (2021): "How to Sell Hard Information," *The Quarterly Journal of Economics*. 7

Alonso, Ricardo and Odilon Câmara (2016a): "Bayesian Persuasion with Heterogeneous Priors," *Journal of Economic Theory*, 165, 672–706. 4

Alonso, Ricardo and Odilon Câmara (2016b): "Political Disagreement and Information in Elections," *Games and Economic Behavior*, 100, 390–412. 8, 17

Asseyer, Andreas and Ran Weksler (2024): "Certification Design with Common Values," *Econometrica*, 92, 651–686. 7

Aumann, Robert J. and Michael Maschler (1995): *Repeated Games with Incomplete Information, with the Collaboration of Richard E. Stearns*, The MIT Press. 8

Ben-Porath, Elchanan, Eddie Dekel, and Barton L. Lipman (2018): "Disclosure and Choice," *The Review of Economic Studies*, 85 (3), 1471–1501. 7

Ben-Porath, Elchanan, Eddie Dekel, and Barton L. Lipman (2021): "Mechanism Design for Acquisition of/Stochastic Evidence," . 7

Bertomeu, Jeremy, Edwige Cheynel, and Davide Cianciaruso (2021): "Strategic Withholding and Imprecision in Asset Measurement," *Journal of Accounting Research*, 59 (5), 1523–1571. 7

Blackwell, David and Meyer A. Girshick (1954): *Theory of Games and Statistical Decisions*, Wiley Publications in Statistics, New York: J. Wiley & Sons. 12

Dasgupta, Sulagna, Ilia Krasikov, and Rohit Lamba (2022): "Hard information design," *working paper*. 7

DeMarzo, Peter M., Ilan Kremer, and Andrzej Skrzypacz (2019): "Test Design and Minimum Standards," *American Economic Review*, 109 (6), 2173–2207. 5, 7, 14, 23

Dworczak, Piotr and Giorgio Martini (2019): "The Simple Economics of Optimal Persuasion," *Journal of Political Economy*, 127 (5), 1993–2048. 4, 8, 10, 17

Dye, Ronald A. (1985): "Disclosure of Nonproprietary Information," *Journal of Accounting Research*, 23 (1), 123–145. 7, 12, 13

Edlin, Aaron S. and Chris Shannon (1998): "Strict Monotonicity in Comparative Statics," *Journal of Economic Theory*, 81 (1), 201–219. 41

Escudé, Matteo (2024): "Covert Learning and Disclosure," *working paper*. 7

Felgenhauer, Mike (2019): "Endogenous Persuasion with Costly Verification," *The Scandinavian Journal of Economics*, 121 (3), 1054–1087. 8

Gentzkow, Matthew and Emir Kamenica (2016): "A Rothschild-Stiglitz Approach to Bayesian Persuasion," *American Economic Review: Papers & Proceedings*, 106 (5), 597–601. 11, 12

Gentzkow, Matthew and Emir Kamenica (2017): "Disclosure of Endogenous Information," *Economic Theory Bulletin*, 5 (1), 47–56. 7

Grossman, Sanford J. (1981): "The Informational Role of Warranties and Private Disclosure about Product Quality," *Journal of Law & Economics*, 24, 461–484. 6, 14

Hagenbach, Jeanne, Frédéric Koessler, and Eduardo Perez-Richet (2014): "Certifiable Pre-Play Communication: Full Disclosure," *Econometrica*, 82 (3), 1093–1131. 7

Jung, Woon-Oh and Young K. Kwon (1988): "Disclosure When the Market Is Unsure of Information Endowment of Managers," *Journal of Accounting Research*, 26 (1), 146–153. 7, 14

Kamenica, Emir (2019): "Bayesian Persuasion and Information Design," *Annual Review of Economics*, 11 (1). 8

Kamenica, Emir and Matthew Gentzkow (2011): "Bayesian Persuasion," *American Economic Review*, 101 (6), 2590–2615. 8

Kartik, Navin, Frances Xu Lee, and Wing Suen (2017): "Investment in Concealable Information by Biased Experts," *The RAND Journal of Economics*, 48 (1), 24–43. 7

Kolotilin, Anton (2018): "Optimal Information Disclosure: A Linear Programming Approach," *Theoretical Economics*, 13 (2), 607–635. 4, 8, 11, 12, 17

Kolotilin, Anton, Timofiy Mylovanov, and Andriy Zapechelnyuk (2022): "Censorship as optimal persuasion," *Theoretical Economics*, 17 (2), 561–585. 8

Kolotilin, Anton, Tymofiy Mylovanov, Andriy Zapechelnyuk, and Ming Li (2017): "Persuasion of a Privately Informed Receiver," *Econometrica*, 85 (6), 1949–1964. 8, 17

Leshno, Moshe, Haim Levy, and Yishay Spector (1997): "A Comment on Rothschild and Stiglitz's "Increasing Risk: I. A Definition"," *Journal of Economic Theory*, 77 (1), 223–228. 12

Lipnowski, Elliot, Doron Ravid, and Denis Shishkin (2021): "Persuasion via Weak Institutions," *working paper available at https://denisshishkin.com/papers/persuasion_via_weak_institutions_2021.pdf*. 17, 18, 33

Lizzeri, Alessandro (1999): "Information Revelation and Certification Intermediaries," *The RAND Journal of Economics*, 30 (2), 214–231. 7

Matthews, Steven and Andrew Postlewaite (1985): "Quality Testing and Disclosure," *RAND Journal of Economics*, 16 (3), 328–340. 7

Milgrom, Paul (2008): "What the Seller Won't Tell You: Persuasion and Disclosure in Markets," *Journal of Economic Perspectives*, 22 (2), 115–131. 6

Milgrom, Paul and John Roberts (1986): "Relying on the Information of Interested Parties," *The RAND Journal of Economics*, 18–32. 6

Milgrom, Paul R. (1981): "Good News and Bad News: Representation Theorems and Applications," *The Bell Journal of Economics*, 12 (2), 380–391. 6, 14

Nguyen, Anh and Teck Yong Tan (2021): "Bayesian persuasion with costly messages," *Journal of Economic Theory*, 193, 105212. 8

Onuchic, Paula (2024): "Advisors with Hidden Motives," *arXiv:2103.07446*. 8

Rothschild, Michael and Joseph E Stiglitz (1970): "Increasing Risk: I. A Definition," *Journal of Economic Theory*, 2 (3), 225–243. 12

Topkis, Donald M (1978): "Minimizing a submodular function on a lattice," *Operations research*, 26 (2), 305–321. 41, 43

Whitmeyer, Mark and Kun Zhang (2022): "Costly Evidence and Discretionary Disclosure," *arXiv preprint arXiv:2208.04922*. 7