# Credible Persuasion[*]

Xiao Lin[†]    Ce Liu[‡]

May 6, 2022

**Abstract**

We propose a new notion of credibility for Bayesian persuasion problems. A disclosure policy is credible if the sender cannot profit from tampering with her messages while keeping the message distribution unchanged. We show that the credibility of a disclosure policy is equivalent to a cyclical monotonicity condition on its induced distribution over states and actions. We also characterize how credibility restricts the Sender's ability to persuade under different payoff structures. In particular, when the sender's payoff is state-independent, all disclosure policies are credible. We apply our results to the market for lemons, and show that no useful information can be credibly disclosed by the seller, even though a seller who can commit to her disclosure policy would perfectly reveal her private information to maximize profit.

[†]Department of Economics, Pennsylvania State University (e-mail: xiao@psu.edu).
[‡]Department of Economics, Michigan State University (e-mail: celiu@msu.edu).

# Contents

# 1   Introduction

When an informed party (Sender; she) discloses information to persuade her audience (Receiver; he), it is in her interest to convey only messages that steer the outcome in her own favor: schools may want to inflate their grading policies to improve their job placement records; similarly, credit rating agencies may publish higher ratings in exchange for future business. Even when the Sender claims to have adopted a disclosure policy, she may still find it difficult to commit to following its prescriptions, since the adherence to such policies is often impossible to monitor. By contrast, what *is* often publicly observable is the final distribution of the Sender's messages: students' grade distributions at many universities are publicly available, and so is the distribution of rating scores from credit rating agencies.

Motivated by this observation, we propose a notion of *credible persuasion*. In contrast to standard Bayesian persuasion, our Sender cannot commit to a disclosure policy; however, to avoid detection, she must keep the final message distribution unchanged when deviating from her disclosure policy. For example, in the context of grade distributions, if a university had announced a disclosure policy that features a certain fraction of A's, B's, and C's, it cannot switch to a distribution that assigns every student an A without being detected. Analogously, if a credit rating agency were to tamper with its rating schemes, any resulting change in the overall distribution of ratings may be detected. Our notion of credibility closely adheres to this definition of detectability: we say that a disclosure policy is credible if given how the Receiver reacts to her messages, the Sender has no profitable deviation to any other disclosure policy that has the same message distribution.

Can the Sender persuade the Receiver by using credible disclosure policies? We find that in many settings, no informative disclosure policy is credible. An important special case where this effect is exhibited is the market for lemons (Akerlof, 1970). Here, we show that the seller of an asset cannot credibly disclose any useful information to the buyer; this effect arises even though the seller benefits from persuasion when she can fully commit to her disclosure policy. Conversely, we also provide conditions for when the Sender is guaranteed to benefit from credible persuasion so that credibility does not entirely eliminate the scope for persuasion. In general, we show that credibility is characterized by a *cyclical monotonicity* condition that is analogous to that studied in decision theory and mechanism design (Rochet, 1987).

To illustrate these ideas, consider the following example. A buyer (Receiver) chooses whether to buy a car from a used car seller (Sender). It is common knowledge that 30% of the cars are of *high* quality and the remaining 70% are of *low* quality. For simplicity, suppose that all cars are sold at an exogenously fixed price.[1] The payoffs in this example are in Table 1.

---

[1]In Section 3 we study a competitive market for lemons with endogenous prices, and emerge with similar findings.

|       | Buy | Not Buy |
|-------|-----|---------|
| High  | 2   | 1       |
| Low   | 2   | 0       |

Seller

|       | Buy | Not Buy |
|-------|-----|---------|
| High  | 1   | 0       |
| Low   | $-1$ | 0      |

Buyer

Table 1: Used Car Example Payoffs

The seller always prefers selling a car, but the buyer is only willing to purchase if and only if he believes its quality is high with at least 0.5 probability. Conditional on a car being sold, the seller obtains the same payoff regardless of its quality; but when a car is not sold, she receives a higher value from retaining a high quality car.

As a benchmark, let us first see what the seller achieves if she could commit to a disclosure policy. We depict the optimal disclosure policy in Figure 1. The policy uses two messages, *pass* and *fail*: all high-quality cars pass, along with 3/7 of the low-quality cars; the remaining 4/7 of the low-quality cars receive a failing grade. Conditional on the car passing, the buyer believes that the car is of high quality with probability 0.5, which is just enough to convince him to make the purchase. If a car fails, the buyer believes that the car is of low quality for sure and will refuse to buy. With this disclosure policy, the buyer expects to see the seller pass 60% of the cars and fail the remaining 40%.

**Prior Distribution**     **Message Distribution**
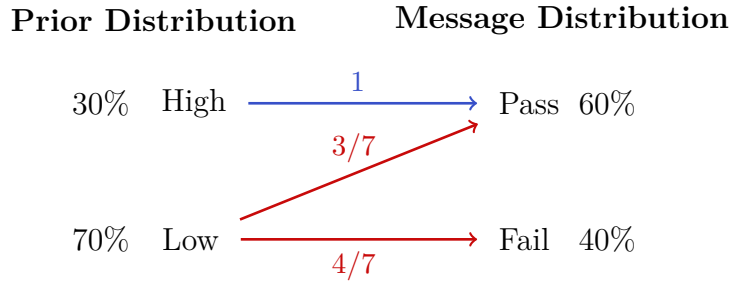


Figure 1: Optimal Commitment Policy

The policy above is optimal for the seller if she can commit to following its prescriptions. But suppose the buyer cannot observe how the seller rates her cars. Instead, the buyer only observes the fraction of cars being passed and failed. In such a setting, the seller can profitably deviate from the above disclosure policy without being detected by the buyer. Specifically, the seller can switch to failing all high-quality cars and passing an equal number of low-quality cars. This disclosure policy, illustrated in Figure 2, induces the same distribution of messages (i.e., 60% pass, 40% fail). Holding fixed the buyer's behavior, this deviation is profitable for the seller because she is still selling the same number of cars but now is able to retain more high-quality cars. Accordingly, we view the optimal full-commitment policy not to be credible:

after having promised to share information according to a disclosure policy, the seller would not find it rational to follow through and would instead profit from an undetectable deviation.

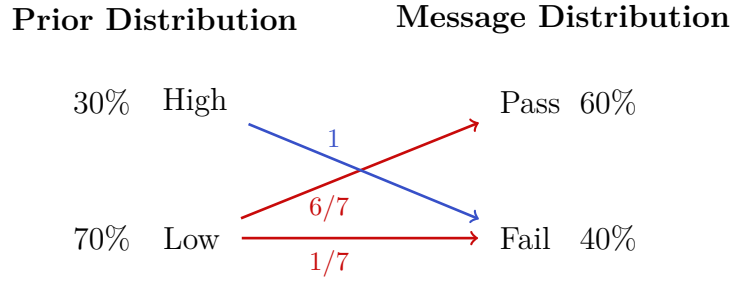**Prior Distribution**     **Message Distribution**



Figure 2: An Undetectable Deviation

More generally, we introduce the following notion of credibility for disclosure policies. Consider a *profile* consisting of the Sender's disclosure policy and the Receiver's strategy (mapping messages to actions). We say that a profile is *Receiver incentive compatible* if the Receiver's strategy best responds to the Sender's disclosure policy—this requirement is standard in Bayesian persuasion problems. We say that a profile is *credible* if, given the Receiver's strategy, the Sender has no profitable deviation to any other disclosure policy that induces the same message distribution. Together, credibility and Receiver incentive compatibility require that conditional on the Sender's message distribution, the Sender and Receiver best respond to each other.[2]

We have just argued that in the used car example, the optimal full-commitment disclosure policy was not credible given the Receiver's best response. Can any car be sold in a profile that is both credible and Receiver incentive compatible? We show that the answer is no. Note that this is what happens when no information is disclosed. In other words, credibility completely shuts down the possibility for useful information transmission.

To see why, suppose towards a contradiction that the buyer purchases a car after observing a message $m_1$ that is sent with positive probability. By Receiver incentive compatibility, the buyer must believe that the car is of high quality with more than 0.5 probability after observing $m_1$. Since $m_1$ is sent with positive probability, the martingale property of beliefs implies that there must be another message $m_2$, also sent with positive probability, that makes the buyer assign less than 0.5 to the car's quality being high. Necessarily, when the buyer observes the message $m_2$, he does not make a purchase. This creates an incentive for the seller to tamper with her disclosure policy: by exchanging some of the good cars being mapped into $m_1$ with an equal number of bad cars being mapped into $m_2$, she can improve her payoff without changing the distribution of messages.

---

[2]Our solution-concept is therefore analogous to a Nash equilibrium condition in which the set of feasible deviations for the Sender is to other disclosure policies that induce the same message distribution.

One may wonder if credibility always shuts down communication entirely. The next example features a setting in which the optimal full-commitment disclosure policy is credible. Consider the disclosure problem faced by a school (Sender) and an employer (Receiver).[3] Just as in the used car example, a student's ability is either *high* with probability 0.3 or *low* with probability 0.7. Payoffs are as shown in Table 2. The employer is willing to hire a student if he believes the student has high ability with at least 0.5 probability. The school would like all its students to be employed, but derives a higher payoff from placing a good student than it does from placing a bad one.

| | Hire | Not Hire | | | Hire | Not Hire |
|------|------|----------|---|------|------|----------|
| High | 2 | 0 | | High | 1 | 0 |
| Low | 1 | 0 | | Low | $-1$ | 0 |
| | School | | | | Employer | |

Table 2: School Example Payoffs

The school's optimal full-commitment disclosure policy is identical to the one in the used car example (Figure 1), and so are the employer's best responses. But unlike the used car example, the school cannot profitably deviate without changing the message distribution.

To see why, note that without changing the message distribution, any deviation must involve passing some low ability students while failing an equal number of high ability students. This would increase the employment of low ability students at the expense of their high ability counterparts, which makes the school worse off. Since the school cannot profit from undetectable deviations, the optimal full-commitment policy is credible. In contrast to the previous example where credibility shuts down all useful communication, the current example shows that credibility sometimes imposes no cost on the Sender relative to persuasion with full commitment.

In the two examples above, credibility has starkly different implications for information transmission. The key difference is that in the used car example, when the car's quality is higher, the Sender has a weaker incentive to trade while the Receiver's incentive to trade is stronger; in the school example, by contrast, both the Sender and Receiver have a stronger incentive to trade as the student's ability increases. Our results formalize this intuition.

Proposition 1 shows that when the Sender and Receiver's preferences have opposite modularities (e.g. when the Sender's payoff is strictly supermodular and the Receiver's payoff is submodular), no useful information can be credibly communicated. When players' preferences share the same modularity, the Sender does not always benefit from credible persuasion

---

[3]See Ostrovsky and Schwarz (2010) for an early study of how schools strategically design their grading policies in a competitive setting.

relative to the no-information benchmark. Proposition 2 and Proposition 3 provide additional conditions that guarantee the Sender does benefit from credible persuasion, as well as conditions under which the optimal full-commitment disclosure policy is credible.

Generalizing further, we use optimal transport theory to characterize credibility using a familiar condition from mechanism design and decision theory—cyclical monotonicity. Theorem 1 shows that for every profile of Sender's disclosure policy and Receiver's strategy, the credibility of the profile is equivalent to a cyclical monotonicity condition on its induced distribution over states and actions. As is illustrated in the examples above, credibility requires that the Sender cannot benefit from any pairwise swapping in the matching of states and actions. The cyclical monotonicity condition generalizes this idea to cyclical swapping: for every sequence of state-action pairs in the support, the sum of the Sender's utility should be lower after the matchings of states and actions in this sequence are permuted. In Section 4.1, we discuss the connection of Theorem 1 to Rochet (1987).

We apply our results to a classic setting of asymmetric information—the market for lemons. It is well-known that market outcome in such a setting may be inefficient due to adverse selection (Akerlof, 1970): despite common knowledge of gain from trade, some cars may not be traded. If the seller can commit to a disclosure policy to persuade the buyers, she can completely solve the market inefficiency by perfectly revealing $\theta$ to the buyers. However, we show that if the buyers can only observe the message distribution of the seller's disclosure policy, but not exactly how these messages are generated, then the seller cannot credibly disclose any useful information to the buyer.

Our paper also offers foundations for studying Bayesian persuasion in a number of settings. One example is when the Sender's payoff is state-independent: in these cases, our results imply that all disclosure policies are credible, so the full-commitment assumption in the Bayesian persuasion approach is nonessential as long as the message distribution is observable. Another example is when the Sender's payoff is supermodular, in which case all monotone disclosure policies are credible.

The rest of the paper is organized as follows: Section 2 introduces our credibility notion as well as the main results. Section 3 considers an application: in the market for lemons with endogenous prices, we show that the seller cannot credibly disclose any useful information to the buyers, even though full disclosure would maximize the seller's profit. Section 4 discusses several aspects of our model: first, it connects our cyclical-monotonicity characterization to Rochet (1987); it then discusses a finite-population approximation of our model, and provides an extensive-form foundation to our partial-commitment formulation. Section 5 concludes. All omitted proofs are in Appendix A. The remainder of this introduction places our contribution within the context of the broader literature.

**Related Literature:** Our work contributes to the study of strategic communication. The Bayesian persuasion model in Kamenica and Gentzkow (2011) studies a Sender who can fully commit to an information structure.[4] By contrast, the cheap-talk approach pioneered by Crawford and Sobel (1982) models a Sender who observes the state privately and, given the Receiver's strategy, chooses an optimal (sequentially rational) message. The partial commitment setting that we model is between these two extremes: here, the Sender can commit to a (marginal) distribution over messages but not the entire information structure.

Our model considers a Sender who can misrepresent her messages as long as the misrepresentation still produces the original message distribution. This contrasts with existing approaches to modeling limited commitment in Bayesian persuasion. One approach, pioneered by Lipnowski, Ravid, and Shishkin (2021) and Min (2021), is to allow the Sender to alter the messages from her chosen test with some fixed probability. Another approach is to consider settings where the Sender can revise her test at a cost. Nguyen and Tan (2021) consider a Sender who can distort the messages from her chosen information structure, whereas Perez-Richet and Skreta (2021) consider a Sender who can falsify the state, or input, of the information structure.

The way that we model the Sender's feasible deviations is closely related to the literature on quota mechanisms, which use message budgets to induce truth-telling; see, for example, Jackson and Sonnenschein (2007), Matsushima, Miyazaki, and Yagi (2010), Rahman (2010), and Frankel (2014). Similar ideas have also been explored in communication games. For example, Chakraborty and Harbaugh (2007) consider multi-issue cheap-talk problems, and study equilibria where the Sender assigns a ranking to each issue.[5] Renault, Solan, and Vieille (2013) study repeated cheap-talk models where only messages and the Receiver's actions are publicly observable. They characterize equilibria in the repeated communication game via a static reporting game where the Sender directly reports her type. The key condition in their characterization requires truthful reporting to be optimal among all reporting strategies that reproduce the true type distribution, which is akin to Rahman (2010)'s characterization of implementable direct mechanisms.[6] Margaria and Smolin (2018) use a different approach to study the case where the Sender's payoff is state-independent, and Meng (2021) provides a unified approach to characterizing the Receiver's optimal value in these repeated cheap-talk models. Kuvalekar, Lipnowski, and Ramos (2021) study a related model where the Receiver is

---

[4]Brocas and Carrillo (2007) and Rayo and Segal (2010) also study optimal disclosure policy in more specific settings.

[5]In such equilibria, the Sender's on-path deviations are essentially restricted to permuting the assignment of rankings to issues. This is similar to quota mechanisms where deviations are restricted to permuting the matching between the Sender's true types and her reported types.

[6]Relatedly, Escobar and Toikka (2013) consider repeated private-value Bayesian games with Markovian states, and show that efficiency can be achieved using a similar quota-like construction.

short-lived, and show that the equilibrium payoffs can be characterized via a static cheap-talk model with capped money burning.

A different strand of the repeated cheap-talk literature studies models where the Receiver can observe feedback about past state realizations. Best and Quigley (2020) considers how coarse feedback of past states can substitute for commitment; Mathevet, Pearce, and Stacchetti (2022) allows for the possibility of non-strategic commitment types; Pei (2020) characterizes when Sender's persistent private information about her own lying cost allows her to achieve her full-commitment payoff.

Finally, our approach to credible persuasion is reminiscent of how Akbarpour and Li (2020) model credible auctions. They study mechanism design problems where the designer's deviations are "safe" so long as they lead to outcomes that are possible when she is acting honestly, and characterize mechanisms that ensure the designer has no safe and profitable deviations. By contrast, we study persuasion problems where the Sender's deviations are undetectable if they do not alter the message distribution, and characterize information structures where the Sender has no profitable and undetectable deviation.

# 2 Model

## 2.1 Setup

We consider an environment with a single Sender ($S$; she) and a single Receiver ($R$; he). Both players' payoffs depend on an unknown state $\theta \in \Theta$ and the Receiver's action $a \in A$. Both $\Theta$ and $A$ are finite sets. The payoff functions are given by $u_S : \Theta \times A \to \mathbb{R}$ and $u_R : \Theta \times A \to \mathbb{R}$. Players hold full-support common prior $\mu_0 \in \Delta(\Theta)$.

Let $M$ be a finite message space that contains $A$. The Sender chooses a disclosure policy, which we henceforth refer to as a "test," to influence the Receiver's action. A test $\lambda \in \Delta(\Theta \times M)$ is a joint distribution of states and messages, so that the marginal distribution of states agrees with the prior; that is, $\lambda_\Theta = \mu_0$.[7] The Receiver chooses an action after observing each message according to a pure strategy $\sigma : M \to A$.[8]

Our interest is in understanding the Sender's incentives to deviate from her test, which depends on the Receiver's strategy. To avoid ambiguity, we refer explicitly to pairs of $(\lambda, \sigma)$—or *profiles*—that consist of a Sender's disclosure policy and a Receiver's strategy. For each

---

[7]For a probability measure $P$ defined on some product space $X \times Y$, we use $P_X$ and $P_Y$ to denote its marginal distribution on $X$ and $Y$, respectively.

[8]We focus on pure strategies to abstract from the Receiver using randomization to deter the Sender's deviations.

profile $(\lambda, \sigma)$, the players' expected payoffs are

$$U_S(\lambda, \sigma) = \sum_{\theta,m} u_S(\theta, \sigma(m))\lambda(\theta, m) \qquad \text{and} \qquad U_R(\lambda, \sigma) = \sum_{\theta,m} u_R(\theta, \sigma(m))\lambda(\theta, m).$$

We consider a setting where the Sender cannot commit to her test, and can deviate to another test so long as it leaves the final message distribution unchanged. This embodies the notion that the distribution of the Sender's messages is observable, even though it may be difficult to observe exactly how these messages are generated. Formally, if $\lambda$ is a test promised by the Sender, let $D(\lambda) \equiv \{\lambda' \in \Delta(\Theta \times M) : \lambda'_\Theta = \mu_0, \lambda'_M = \lambda_M\}$ denote the set of tests that induce the same distribution of messages as $\lambda$: these test are indistinguishable from $\lambda$ from the Receiver's perspective. Our credibility notion requires that conditioning on how the Receiver responds to the Sender's messages, no deviation in $D(\lambda)$ can be profitable for the Sender.

**Definition 1.** *A profile $(\lambda, \sigma)$ is **credible** if*

$$\lambda \in \underset{\lambda' \in D(\lambda)}{\arg\max} \sum_{\theta,m} u_S(\theta, \sigma(m)) \, \lambda'(\theta, m) \tag{1}$$

Moreover, the Receiver's strategy is required to be a best response to the Sender's chosen test.

**Definition 2.** *A profile $(\lambda, \sigma)$ is **Receiver Incentive Compatible** (R-IC) if*

$$\sigma \in \underset{\sigma':M \to A}{\arg\max} \sum_{\theta,m} u_R(\theta, \sigma'(m)) \, \lambda(\theta, m) \tag{2}$$

Together, credibility and R-IC ensure that conditioning on the message distribution of the Sender's test, both the Sender and the Receiver best respond to each other.

An immediate observation is that there always exists a profile $(\lambda, \sigma)$ that is both credible and R-IC. This is the profile of a completely uninformative test and a Receiver strategy that takes the ex ante optimal action after every message. Given the test, the Receiver is taking a best response and given the Receiver's strategy, the Sender has no incentive to deviate to any other test that induces the same message distribution.

Our credibility notion is motivated by the observability of the Sender's message distribution, which is modeled as a restriction on the Sender's feasible deviations. Below we first discuss the observability of message distribution in light of the population interpretation of Bayesian persuasion models; we then connect this observability to the partial commitment approach we adopt in our model.

The observability of the message distribution is best understood through a population interpretation of persuasion models,[9] where there are a continuum of objects with types distributed according to $\mu_0 \in \Delta(\Theta)$. The Sender's test $\lambda$ assigns each object a message based on its type, which generates a message distribution $\lambda_M$. Working with a continuum population affords us a cleaner exposition by abstracting from sampling variation. In Section 4.2, we consider a finite approximation where the Sender privately observes $N$ i.i.d. samples from $\mu_0 \in \Delta(\Theta)$, and assigns each realization a message $m \in M$ subject to quotas on message frequencies; the Receiver then chooses an action after observing the Sender's message. We show that credible and R-IC profiles in our continuum model are approximated by those in the finite-sample model when the sample size $N$ becomes large.

The credibility notion in Definition 1 restricts the Sender's feasible deviations by requiring them to preserve the original message distribution. To formalize the connection between the observability of message distribution and the Sender's partial commitment, in Section 4.3 we present an extensive-form game in which the Sender is permitted to deviate to any test, and the Receiver observes the message distribution generated by the chosen test. As we show therein, if the Sender chooses a test that induces a different message distribution, the Receiver can "punish" the Sender by assuming that the chosen test is uninformative. If one focuses on profiles where the Sender is weakly better off than disclosing no information (as we do in this paper), this analysis provides a justification for our focus on deviations that generate the same distribution of messages as the equilibrium test.

## 2.2 Stable Outcome Distributions

We characterize credible and receiver incentive compatible profiles through the induced probability distribution of states and actions. Formally, an *outcome distribution* is a distribution $\pi \in \Delta(\Theta \times A)$ that satisfies $\pi_\Theta = \mu_0$: this is a consistency requirement that stipulates that the marginal distribution of states must conform to the prior. We say an outcome distribution $\pi$ is induced by a profile $(\lambda, \sigma)$ if for every $(\theta, a) \in \Theta \times A$, $\pi(\theta, a) = \lambda(\theta, \sigma^{-1}(a))$, where $\sigma^{-1}$ is the inverse mapping of $\sigma$. We are interested in characterizing outcome distributions that can be induced by profiles that are both credible and R-IC, and refer to such outcome distributions as stable.

**Definition 3.** *An outcome distribution $\pi \in \Delta(\Theta \times A)$ is **stable** if it is induced by a profile $(\lambda, \sigma)$ that is both credible and R-IC.*

Our first result characterizes stable outcome distributions.

---

[9] For a more detailed discussion of various interpretations of Bayesian persuasion models, see e.g. Section 2.2 of Kamenica (2019).

**Theorem 1.** *An outcome distribution $\pi \in \Delta(\Theta \times A)$ is stable if and only if:*

1. *$\pi$ is $u_R-$obedient: for each $a \in A$ such that $\pi(\Theta, a) > 0$,*

$$\sum_{\theta \in \Theta} \pi(\theta, a)\, u_R(\theta, a) \geq \sum_{\theta \in \Theta} \pi(\theta, a)\, u_R(\theta, a') \ \ for\ all\ \ a' \in A.$$

2. *$\pi$ is $u_S-$cyclically monotone: for each sequence $(\theta_1, a_1), \ldots, (\theta_n, a_n) \in supp(\pi)$ and $a_{n+1} \equiv a_1$,*

$$\sum_{i=1}^{n} u_S(\theta_i, a_i) \geq \sum_{i=1}^{n} u_S(\theta_i, a_{i+1});$$

The first condition is the standard obedience constraint (Bergemann and Morris, 2016; Taneva, 2019), which specifies that the Receiver finds it incentive compatible to follow the recommended action given the belief that she forms when receiving that recommendation. The second condition, namely $u_S$-cyclical monotonicity, is the new constraint that maps directly to our notion of credibility. Below, we describe this condition and explain why it is both necessary and sufficient for credibility.

To understand the cyclical monotonicity condition, consider an outcome distribution $\pi$ and a sequence $(\theta_i, a_i)_{i=1}^n$ in the support of $\pi$. A "cyclical" deviation in this case consists of subtracting $\varepsilon$ mass from $(\theta_i, a_i)$ while adding it to $(\theta_i, a_{i+1})$ for each $i = 1, \ldots, n$, where $a_{n+1} \equiv a_1$. Each step of this cyclical deviation changes the Sender's payoff by $\varepsilon\big[u_S(\theta_i, a_{i+1}) - u_S(\theta_i, a_i)\big]$, so the total change in the Sender's payoff is

$$\varepsilon\Big[\sum_{i=1}^{n} u_S(\theta_i, a_{i+1}) - \sum_{i=1}^{n} u_S(\theta_i, a_i)\Big].$$

The cyclical monotonicity condition requires that the Sender can find no profitable cyclical deviations.

To see why cyclical monotonicity is necessary, observe that cyclical deviations do not change the distribution of recommended actions. Therefore, any such deviation could not be detected solely on the basis of the distribution of messages. Because we require that such undetectable deviations are not profitable, this implies the cyclical monotonicity condition above.

For sufficiency, the intuition is that any deviation that keeps the marginal distribution on messages unchanged can either be expressed as or approximated by cyclical deviations. To see the idea, let us look at the graphical representation of an outcome distribution $\pi$ in the left panel of Figure 3. In this example, $\Theta = \{\square, \bigcirc, \triangle\}$ and $A = \{x, y, z\}$. Each $\square$, $\bigcirc$, and $\triangle$ in the graph is associated with 10% probability mass. The prior belief assigns 20% to $\triangle$, which
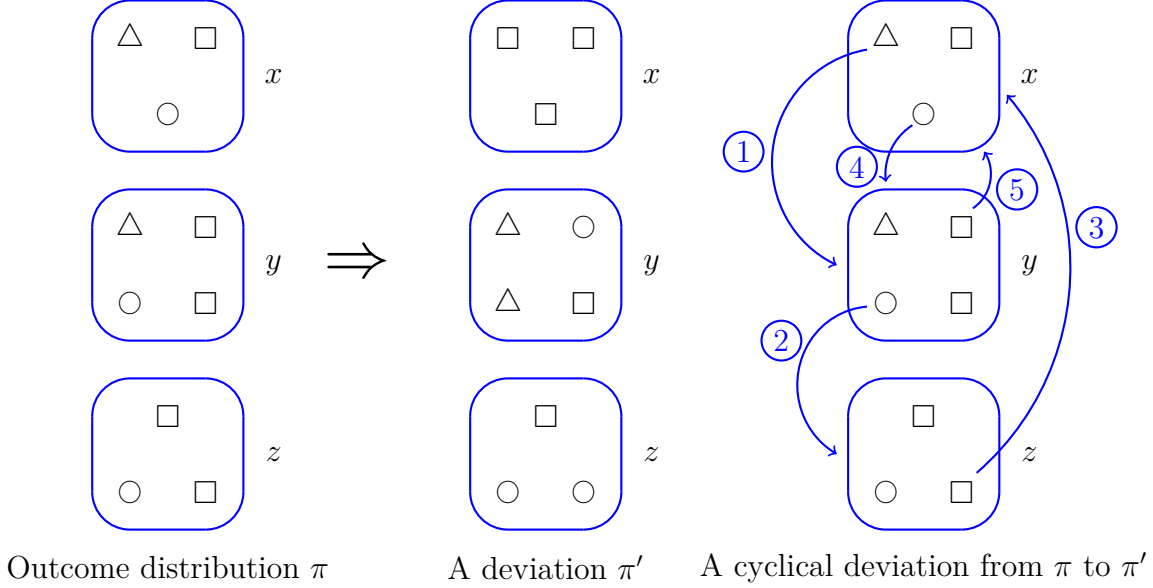
Figure 3: Cyclical Deviation

is represented by the presence of two $\triangle$'s; similarly, the prior assigns 30% to $\bigcirc$ and 50% to $\square$. The pairing between states and actions pins down the outcome distribution, as well as its induced distribution of actions. For example, $\pi(\triangle, x) = 10\%$ and $\pi(\Theta, y) = 40\%$.

The middle panel depicts a possible deviation $\pi'$ that maintains the same distribution of actions. In particular, the matchings among states and actions are permuted, but the number of shapes matched to each action remains the same. Our credibility notion requires that no such deviation can be profitable. The right panel illustrates how this deviation can be expressed as a 5-step cyclical deviation.

In fact, it is not hard to see from the graph that every deviation that involves moving integer numbers of $\square$, $\bigcirc$, or $\triangle$ is essentially a permutation of the locations of the shapes, which can be written as a disjoint composition of cycles. It is therefore sufficient to check for deviations that form a cycle. However, an implicit assumption of the graphical representation is that both the original outcome distribution and the deviation contain only rational probabilities; otherwise, we cannot find an $N$ large enough so that every joint probability can be represented as integer numbers of shapes. In this case, however, one can approximate irrational outcome distributions with rational ones, and then apply the same argument outlined above. In Appendix A.1, we prove Theorem 1 by directly citing Kantorovich duality from optimal transport theory.

The final remark is that in the cyclical monotonicity condition, we can restrict attention to those sequences with length $n \leq \max\{|\Theta|, |M|\}$. As an illustration, the right panel of Figure 3 represents a cyclical deviation of length 5, which is greater than both the number of

11

states and the number of actions. However, we can split this cycle into two shorter cyclical deviations, one comprising ①②③ and the other made up of ④⑤. If the deviation of length 5 is profitable, one of these two shorter sequences has to be profitable. Theorem 1 therefore reduces the problem to checking only a finite number of deviations.

## 2.3   The Case of Additively Separable Payoffs

If $u_S(\theta, a)$ is additively separable in $\theta$ and $a$, then $u_S-$cyclical monotonicity is automatically satisfied. So we have the following observation.

**Observation 1.** *If $u_S(\theta, a) = v(\theta) + w(a)$ for some $v : \Theta \to \mathbb{R}$ and $w : A \to \mathbb{R}$, then every outcome distribution that satisfies $u_R$-obedience is stable.*

Therefore, in this case, there is no gap between what is achievable by a sender who can fully commit to a test relative to a sender who can only partially commit to a distribution of messages. This observation is relevant to a special and widely studied case of additively separable preferences, namely that in which the Sender has state-independent payoffs. State-independent payoffs feature in many analyses of communication and persuasion (e.g. Chakraborty and Harbaugh, 2010; Alonso and Câmara, 2016; Lipnowski and Ravid, 2020; Lipnowski, Ravid, and Shishkin, 2021; Gitmez and Molavi, 2022). Our analysis suggests that for these settings, the Sender can persuade with full commitment power even without committing to a test, merely by making public (and committing to) her distribution of messages.

## 2.4   When is Credibility Restrictive?

When the state and action interact in the Sender's payoff, credibility limits the Sender's choice of tests. The goal of this section is to understand how these limits can restrict the Sender's ability to persuade the Receiver. As benchmarks, we will often draw comparisons to what the Sender can achieve when she can fully commit to her disclosure policy, as well as what is achievable when all or no information is disclosed. We say an outcome distribution $\pi^*$ is an *optimal full-commitment outcome* if it maximizes the Sender's payoff among outcome distributions that satisfy $u_R$-obedience. An outcome distribution $\hat{\pi}$ is a *fully revealing outcome* if the Receiver always chooses a best response to every state; that is,

$$a \in \arg\max_{a' \in A} u_R(\theta, a') \ \text{ for every } \ (\theta, a) \in \text{supp}(\hat{\pi}).$$

Finally, an outcome distribution $\pi^\circ$ is a *no-information outcome* if the Receiver always chooses the same action that best responds to the prior belief $\mu_0$; in other words, there exists

$$a^* \in \arg\max_{a \in A} \sum_{\theta \in \Theta} \mu_0(\theta) u_R(\theta, a) \ \text{ such that } \ \pi_A^\circ(a^*) = 1.$$

We say the Sender *benefits from persuasion* if an optimal full-commitment outcome gives the Sender a higher payoff than every no-information outcome. Similarly, we say the Sender *benefits from credible persuasion* if there exists a stable outcome distribution that gives the Sender a strictly higher payoff than every no-information outcome.

We make a few assumptions for ease of exposition. First, suppose every $a \in A$ is a best response to some belief $\mu \in \Delta(\Theta)$ for the receiver. This assumption is without loss of generality, since an action that is never a best response would never be played by the Receiver in any R-IC profile, and can be removed from the action set $A$ without changing results in this paper. Second, suppose there exists no distinct $a, a' \in A$ such that $u_R(\theta, a) = u_R(\theta, a')$ for all $\theta \in \Theta$; in other words, from the Receiver's perspective, there are no duplicate actions. This second assumption is *not* without loss, but greatly simplifies the statements of Proposition 1 and Proposition 2.

**Modular Preferences:** In the examples in Section 1, we see that whether the Sender can credibly persuade the Receiver depends crucially on the alignment of their marginal incentives to trade. To understand this logic more generally, we assume that $\Theta$ and $A$ are totally ordered sets, which without loss of generality can be assumed to be subsets of $\mathbb{R}$. Recall that a payoff function $u : \Theta \times A \to \mathbb{R}$ is *supermodular* if for all $\theta > \theta'$ and $a > a'$, we have

$$u(\theta, a) + u(\theta', a') \geq u(\theta, a') + u(\theta', a).$$

and *submodular* if

$$u(\theta, a) + u(\theta', a') \leq u(\theta, a') + u(\theta', a).$$

Furthermore, the function is strictly supermodular or strictly submodular if the inequalities above are strict for $\theta > \theta'$ and $a > a'$.

The modularity of players' payoff functions captures how the marginal utility from higher actions varies with the state. This generalizes the marginal incentive to trade in the examples in Section 1: intuitively, the Sender and the Receiver have aligned marginal incentives when both players' payoff functions share the same modularity, and opposed marginal incentives when their payoff functions have opposite modularities. To fix ideas, we will assume that the Sender's payoff is supermodular and vary the modularity of the Receiver's payoff.

We now introduce a lemma that simplifies the $u_S$-cyclical monotonicity condition in Theorem 1 when the Sender's payoff is supermodular. Say that an outcome distribution $\pi \in \Delta(\Theta \times A)$ is *comonotone* if for all $(\theta, a)$, $(\theta', a') \in \text{supp}(\pi)$ satisfying $\theta < \theta'$, we have $a \leq a'$. Comonotonicity requires that the states and the Receiver's actions are positive-assortatively matched in the outcome distribution. The following lemma, whose variant appears in Rochet (1987), shows that $u_S$-cyclical monotonicity reduces to comonotonicity when the Sender's preference is supermodular.

**Lemma 1.** *If $u_S$ is supermodular, then every comonotone outcome distribution is $u_S$-cyclically monotone. Furthermore, if $u_S$ is strictly supermodular, then every $u_S$-cyclically monotone outcome distribution is also comonotone.*

Combined with Theorem 1, Lemma 1 implies that when the Sender's preference is strictly supermodular, the credibility of a profile $(\lambda, \sigma)$ is equivalent to the comonotonicity of its induced outcome distribution. Comonotone outcome distributions have attracted much attention in the persuasion literature in part due to their simplicity and ease of implementation; for example, see Dworczak and Martini (2019), Goldstein and Leitner (2018), Mensch (2021), Ivanov (2020), Kolotilin (2018), and Kolotilin and Li (2020). Our credibility notion provides an additional motive for focusing on monotone information structures.

**When Credibility Shuts Down Communication:** The next result generalizes the used-car example in Section 1.

**Proposition 1.** *If $u_S$ is strictly supermodular and $u_R$ is submodular, then every stable outcome distribution is a no-information outcome.*

Proposition 1 says that when the players have opposed marginal incentives, credibility considerations completely shut down information transmission. The logic generalizes what we saw in the example: if two distinct messages resulted in different actions from the Receiver, the Sender and Receiver would have diametrically opposed preferences regarding which action to induce in which state. Therefore, if a profile satisfies R-IC and is even partially informative, the Sender would have a profitable deviation to another test that swaps states and induces the same marginal distribution of messages.

**When the Sender Benefits from Credible Persuasion:** In light of the school example in Section 1, one might expect credibility to not limit the Sender's ability to persuade when her marginal incentives are aligned with the Receiver's. However, this is false without imposing

additional assumptions. For an illustration, consider the following example, in which both the Sender and Rceiver have supermodular payoffs. The Sender benefits from persuasion when she can fully commit to her disclosure policy, but no stable outcome distribution can give her a higher payoff than the best no-information outcome.



| $u_S(\theta,a)$ | $a_1$ | $a_2$ | $a_3$ | $a_4$ |
|---|---|---|---|---|
| $\theta = H$ | $-1$ | $0.75$ | $1$ | $0$ |
| $\theta = L$ | $0$ | $0.75$ | $0.5$ | $-1$ |

| $u_R(\theta,a)$ | $a_1$ | $a_2$ | $a_3$ | $a_4$ |
|---|---|---|---|---|
| $\theta = H$ | $0$ | $0.6$ | $0.8$ | $1$ |
| $\theta = L$ | $1$ | $0.8$ | $0.6$ | $0$ |

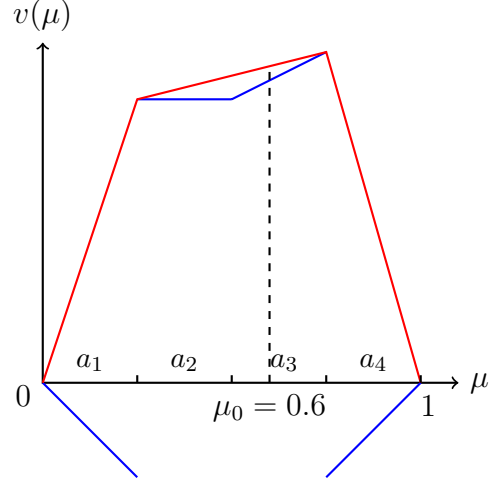Table 3: Sender and Receiver's payoffs          Figure 4: Concavification

**Example 1.** *Suppose $\Theta = \{H, L\}$ with prior $\mu_0 = P(\theta = H) = 0.6$ and $A = \{a_1, a_2, a_3, a_4\}$. The Sender and Receiver's payoffs are as given in Table 3. Note that both players' payoffs are strictly supermodular. The Receiver's best response is $a_1$ when $\mu_0 \in [0, 0.25)$, $a_2$ when $\mu_0 \in [0.25, 0.5)$, $a_3$ when $\mu_0 \in [0.5, 0.75)$, and $a_4$ when $\mu_0 \in [0.75, 1]$; this leads to the Sender's indirect utility function (blue) and its concave envelope (red) depicted in Figure 4. From Kamenica and Gentzkow (2011), the red line represents the Sender's optimal value under full commitment. It is clear that at $\mu_0 = 0.6$, the Sender strictly benefits from persuasion if she can fully commit to her test.*

*However, no stable outcome distribution can make the Sender better off than the no-information outcome. To see why, consider any stable outcome distribution that has at least two actions in the support. By Lemma 1, at most one of these actions can be matched with more than one state, for otherwise the outcome distribution would not be comonotone. So at most one of the actions in the support of the outcome distribution can be induced by interior posterior beliefs. However, it is clear that in order for the Sender to benefit from persuasion, she must induce both $a_2$ and $a_3$, both of which can only happen when the Receiver holds interior posterior beliefs. As a result, no stable outcome distribution can make the Sender better off.*

Example 1 above shows that besides the co-modularity of preferences, additional conditions

are needed in order to ensure the Sender can benefit from credible persuasion. Proposition 2 below offers several such conditions.

Let $\bar{a} \equiv \max A$ and $\underline{a} \equiv \min A$ denote the highest and lowest Receiver actions, and let $\bar{\theta} \equiv \max \Theta$ and $\underline{\theta} \equiv \min \Theta$ denote the highest and lowest states.

**Proposition 2.** *Suppose both $u_S$ and $u_R$ are supermodular.*

1. *If the highest action is dominant for the Sender, that is, if $u_S(\theta, \bar{a}) > u_S(\theta, a)$ for all $\theta$ and $a \neq \bar{a}$, then for generic priors,[10] the Sender benefits from credible persuasion as long as she benefits from persuasion.*

2. *If the Sender favors extreme actions in extreme states, that is, if $u_S(\bar{\theta}, \bar{a}) > u_S(\bar{\theta}, a)$ for all $a \neq \bar{a}$ and $u_S(\underline{\theta}, \underline{a}) > u_S(\underline{\theta}, a)$ for all $a \neq \underline{a}$, then for generic priors, the Sender benefits from credible persuasion.*

3. *If the Sender is strictly better off from a fully revealing outcome than from every no-information outcome, then the Sender benefits from credible persuasion.*

The first condition in Proposition 2 is satisfied in settings like the school example, where the school would always want to place a student regardless of the student's ability. The second condition is applicable in environments where both parties have agreement on extreme states. For example, both doctors and patients favor aggressive treatment if the patient's condition is severe, and both favor no treatment if the patient is healthy, but they might disagree in intermediate cases. Lastly, a special case of the third condition is quadratic loss preferences as commonly used in models of communication (e.g. Crawford and Sobel, 1982).

The first two parts of Proposition 2 are based on belief splitting. Let us briefly describe the proof under the first condition; the proof for the second part follows similar arguments. Note that if $\bar{a}$ is a dominant action for the Sender, and the Sender can benefit from persuasion (under full commitment), then $\bar{a}$ must not already be a best response for the Receiver under the prior $\mu_0$. The Sender can then split the prior into a point mass posterior $\delta_{\bar{\theta}}$ and some other posterior $\tilde{\mu}$ that is close to $\mu_0$. At $\delta_{\bar{\theta}}$, the Receiver is induced to choose $\bar{a}$ since his payoff is supermodular. In addition, for generic priors the Receiver's best response to $\tilde{\mu}$ remains the same as his best response to $\mu_0$. The Sender benefits from this belief-splitting since the same action is still played most of the time, but in addition her favorite action is now played with positive probability. Moreover, the resulting outcome distribution matches higher states with higher actions, so it is stable due to the supermodularity of $u_S$ and Lemma 1.

The intuition for the third part of Proposition 2 is straightforward to see when the Sender's payoff is strictly supermodular. Consider $(\theta, a)$ and $(\theta', a')$ in the support of a fully revealing

---

[10]Formally, by generic we mean a set of priors $T \subset \Delta(\Theta)$ with the same Lebesgue measure as $\Delta(\Theta)$.

outcome distribution $\pi$, so $a$ and $a'$ best respond to $\theta$ and $\theta'$, respectively. From Topkis (2011), it follows that $a \geq a'$ if $\theta > \theta'$. Therefore, $\pi$ is comonotone and satisfies $u_S$-cyclical monotonicity by Lemma 1. By construction, $\pi$ also satisfies obedience, so $\pi$ is stable by Theorem 1.

**When Credibility Imposes No Cost to the Sender:** In Observation 1, we see that when the Sender's payoff is additively separable, credibility does not restrict the set of stable outcomes. Proposition 3 below provides a condition which guarantees that credibility imposes no loss on the Sender's optimal value, even when credibility does restrict the set of stable outcomes.

**Proposition 3.** *Suppose $|A| = 2$. If both $u_S$ and $u_R$ are supermodular, then at least one optimal full-commitment outcome is stable; if in addition $u_S$ is strictly supermodular, then every optimal full-commitment outcome is stable.*

Proposition 3 says that in setting where both players have supermodular payoffs and the Receiver faces a binary decision, such as "accept" or "reject", then credibility imposes no cost to the Sender. This result follows from combining our Theorem 1 and Lemma 1 with Theorem 1 in Mensch (2021). He shows that under the assumptions in our Proposition 3, there exists a optimal full-commitment outcome that is comonotone. The intuition is that for any outcome distribution $\pi$ that is $u_R$-obedient but not comonotone, the Sender can weakly improve her payoff by swapping the non-comonotone pairs in the support of $\pi$, so that they become matched assortatively. Such swapping also benefits the Receiver due to the supermodularity of $u_R$, so $u_R$-obedience remains satisfied. As a result, the Sender can always transform a non-comonotone outcome distribution into one that is comonotone without violating $u_R$-obedience, while weakly improving her own payoff. Therefore, there must be an optimal full-commitment outcome that is comonotone, which is also stable by Theorem 1 and Lemma 1.

# 3 Application: The Market for Lemons

A classic insight from Akerlof (1970) is that in markets with asymmetric information, adverse selection can lead to dramatic efficiency loss. In practice, buyers and sellers often rely on warranty or third-party certification to overcome this inefficiency. A seemingly more direct solution to their predicament is for the seller to fully reveal her private information, so that there is no information asymmetry between players. In this section, however, we show that this apparently easy fix to the adverse selection problem relies on unrealistic assumptions on

the seller's ability to commit. Indeed, we show that any information disclosure that improves efficiency cannot be credible.

To fix ideas, we adapt the formulation in Mas-Colell, Whinston, and Green (1995) and consider a seller who values an asset she owns (say, a car) at $\theta \in \Theta \subset [0,1]$; two buyers (1 and 2) both value the car at $v(\theta)$ which is increasing in $\theta$. Buyers share a common prior belief $\mu_0 \in \Delta(\Theta)$. We assume $v(\theta) > \theta$ for all $\theta \in \Theta$ so there is common knowledge of gain from trade. Moreover, we assume $E_{\mu_0}[v(\theta)] < 1$, so that without information disclosure, some cars will not be traded due to adverse selection. Below we first describe the base game without information disclosure, then augment the base game to allow the seller to choose a test to influence the buyers' belief.

**The Base Game $G$:** The seller and the buyers move simultaneously. The seller learns her value and chooses an ask price $a_s \in A_S = [0, v(1)]$; each buyer $i = 1, 2$ chooses a bid $b_i \in A_i = [0, v(1)]$. If the ask price is lower than or equal to the highest bid, the car is sold at the highest bid to the winning buyer, and ties are broken evenly. If the ask price is higher than the highest bid, the seller keeps the car and receives the reserve value $\theta$, while both buyers get 0. More formally, the seller's payoff function is

$$u_S(\theta, a_S, b_1, b_2) = \begin{cases} \max\{b_1, b_2\} & \text{if } a_S \leq \max\{b_1, b_2\} \\ \theta & \text{if } a_S > \max\{b_1, b_2\} \end{cases}$$

and buyer $i$'s payoff is

$$u_i(\theta, a_S, b_1, b_2) = \begin{cases} v(\theta) - b_i & \text{if } b_i > b_{-i} \text{ and } b_i \geq a_S \\ \frac{1}{2}[v(\theta) - b_i] & \text{if } b_i = b_{-i} \text{ and } b_i \geq a_S \\ 0 & \text{otherwise.} \end{cases}$$

**The Game with Disclosure:** Before the base game is played, the seller chooses a test $\lambda$ to publicly disclose information to the buyers.[11] Together the test $\lambda$ and the base game $G$ defines a Bayesian game $\langle G, \lambda \rangle$. Every message $m$ from the test $\lambda$ induces a posterior belief $\mu_m \equiv \lambda(\cdot|m) \in \Delta(\Theta)$ for the buyers. The buyers $i = 1, 2$ choose their respective bids $\beta_i(m)$, while the seller choose an ask price $\alpha_S(\theta, m)$. We restrict attention to Bayesian Nash equilibria where the seller plays her *weakly dominant strategy* $\alpha_S(\theta, m) = \theta$. As we show in

---

[11]In our setting, $\lambda$ determines only the buyers' information structure, and the seller is perfectly informed about $\theta$. That is, the seller cannot prevent herself from learning the true quality of the car. This differs from Kartik and Zhong (2019), who fully characterize payoffs in the market for lemons under all possible information structures.

Lemma 5, such equilibria exist in $\langle G, \lambda \rangle$ for every $\lambda$. These equilibria also give rise to the familiar fixed-point characterization of equilibrium price: buyers' bids satisfy

$$\max \left\{ \beta_1(m), \beta_2(m) \right\} = E_{\mu_m} \left[ v(\theta) | \theta \leq \max \left\{ \beta_1(m), \beta_2(m) \right\} \right].$$

The trading game above differs from the Sender-Receiver setting in Section 2 in two ways: first, the Sender in the current setting publicly discloses information to multiple Receivers; second, in addition to the Receivers, the Sender also chooses an action (ask price) after observing the realization of the test. Nevertheless, the notion of stable outcome distribution extends to the current setting. In particular, the credibility notion is based on the same idea that the Sender cannot profitably deviate to a different test without changing the message distribution. The Receiver Incentive Compatible (R-IC) condition, meanwhile, is replaced by a new IC condition that asks both the Sender and Receivers to play according to a Bayesian Nash equilibrium in $\langle G, \lambda \rangle$. As mentioned above, in the market for lemons we will focus on a special class of Bayesian Nash equilibria where the seller plays her *weakly dominant strategy* $\alpha_S(\theta, m) = \theta$ in the game $\langle G, \lambda \rangle$. We will call such profiles $(\lambda, \sigma)$ WD-IC to distinguish from the weaker IC requirement. The formal discussion of our credibility notion in this multiple-Receiver setting is notationally cumbersome, and is deferred to Appendix B.

Next we state our result, discuss its implications, and provide intuition for its proof. As a benchmark, fix an arbitrary message $m_0 \in M$, and let $\lambda_0 \equiv \mu_0 \times \delta_{m_0}$ be a null information structure. Let $R_0$ denote the supremum of the seller's payoffs among profiles $(\lambda_0, \sigma)$ that are WD-IC, so $R_0$ represents the highest equilibrium payoff the seller can achieve when providing no information.

**Proposition 4.** *Under every credible and WD-IC profile, the seller's payoff is no more than* $R_0$.

Proposition 4 implies that any information that can be credibly disclosed is not going to improve the seller's payoff compared to the no-information benchmark. This is in sharp contrast to the full-commitment case, where the seller would like to fully reveal the car's quality, and all car types $\theta$ are sold at $v(\theta)$, which would allow the seller to capture all surplus from trade.

Let us describe the intuition behind the proof for Proposition 4. For each message $m$ from the seller's test $\lambda$, let $\Theta(m)$ denote the support of the buyer's posterior belief after observing $m$. A key step in proving Proposition 4 is to show that there exists a common trading threshold $\tau$ such that for each message $m$, a car of quality $\theta \in \Theta(m)$ is traded if and only if $\theta \leq \tau$. To see why, suppose towards a contradiction that the trading threshold in message $m$ is higher than the threshold in another message $m'$. We show in the proof that the

seller would then have a profitable deviation by swapping some of the cars slightly below the higher threshold in message $m$ with an equal amount of cars from $m'$ that are slightly above its lower threshold.[12] Because this deviation does not change the seller's message distribution, it is also undetectable. Therefore, credibility demands a common threshold $\tau$ that applies across messages. Given this common threshold $\tau$, we then apply Tarski's fixed-point theorem to show that when no information is disclosed, there is an equilibrium that features a higher trading threshold $\tau' \geq \tau$. Since a higher threshold means more cars are being traded, which in turn increases the seller's payoff, the seller's payoff under every stable outcome is therefore weakly dominated by her payoff from a no-information outcome, and this proves our result.

# 4  Discussion

## 4.1  Relationship to Rochet (1987)

The $u_S$-cyclical monotonicity condition in our characterization closely resembles the cyclical monotonicity condition for implementing transfers in Rochet (1987). The reader might wonder why cyclical monotonicity arises in our setting despite the lack of transfers. The connection is best summarized by the following three equivalent conditions from optimal transport theory (see, for example, Theorem 5.10 of Villani (2008)).

**Kantorovich Duality.** *Suppose $X$ and $Y$ are both finite sets, and $u : X \times Y \to \mathbb{R}$ is a real-valued function. Let $\mu$ be a probability measure on $X$ and $\nu$ be a probability measure on $Y$, and $\Pi(\mu, \nu)$ be the set of probability measures on $X \times Y$ such that the marginals on $X$ and $Y$ are $\mu$ and $\nu$, respectively. Then for any $\pi^* \in \Pi(\mu, \nu)$, the following three statements are equivalent:*

1. *$\pi^* \in \arg\max_{\pi \in \Pi(\mu,\nu)} \sum_{x,y} \pi(x,y) u(x,y)$;*

2. *$\pi^*$ is $u$-cyclically monotone. That is, for any $n$ and $(x_1, y_1), ..., (x_n, y_n) \in supp(\pi^*)$,*

$$\sum_{i=1}^{n} u(x_i, y_i) \geq \sum_{i=1}^{n} u(x_i, y_{i+1})$$

3. *There exists $\psi : Y \to \mathbb{R}$ such that for any $(x, y) \in supp(\pi^*)$ and any $y' \in Y$,[13]*

$$u(x, y) - \psi(y) \geq u(x, y') - \psi(y').$$

---

[12]This deviation is profitable because it allows the seller to replace the higher-quality cars traded in $m$ with the lower-quality, untraded cars in $m'$. After this swapping, the lower-quality cars are now sold at the price for the higher-quality cars in $m$, while the higher-quality cars are now retained by the seller in $m'$.

[13]This statement can also be equivalently written as: there exists $\phi : X \to \mathbb{R}$ and $\psi : Y \to \mathbb{R}$, such that $\phi(x) + \psi(y) \geq u(x,y)$ for all $x$ and $y$, with equality for $(x,y)$ in the support of $\pi^*$.

Our Theorem 1 builds on the equivalence between 1 and 2 in the Kantorovich duality theorem above to show the equivalence between credibility and $u_S$-cyclical monotonicity.

Rochet (1987)'s classic result on implementation with transfers follows from the equivalence between 2 and 3. To see this, consider a principal-agent problem where the agent's private type space is $\Theta$ with full-support prior $\mu_0$, and the principal's action space is $A$. The agent's payoff is $u(\theta, a) - t$, where $t$ is the transfer she makes to the principle. Given an allocation rule $q : \Theta \to A$, let $v_q(\theta, \theta') \equiv u(\theta, q(\theta'))$ denote the payoff that a type-$\theta$ agent obtains from the allocation intended for type $\theta'$. Let $X = Y = \Theta$ and $\mu = \nu = \mu_0$ in the Kontorovich Duality theorem above, and consider the distribution $\pi^* \in \Pi(\mu, \nu)$ defined by

$$
\pi^*(\theta, \theta') = \begin{cases} \mu_0(\theta) & \text{if } \theta = \theta' \\ 0 & \text{otherwise} \end{cases}
$$

By the equivalence of 2 and 3 in the Kantorovich duality theorem, $\pi^*$ is $v_q$-cyclically monotone if and only if there exists $\psi : \Theta \to \mathbb{R}$ such that for all $\theta, \theta' \in \Theta$, $v_q(\theta, \theta) - \psi(\theta) \geq v_q(\theta, \theta') - \psi(\theta')$. That is,

$$
u(\theta, q(\theta)) - \psi(\theta) \geq u(\theta, q(\theta')) - \psi(\theta'),
$$

so the allocation rule $q$ can be implemented by the transfer rule $\psi : \Theta \to \mathbb{R}$. The $v_q$-cyclical monotonicity condition says that for every sequence $\theta_1, ..., \theta_n \in \Theta$ with $\theta_{n+1} \equiv \theta_1$,

$$
\sum_{i=1}^{n} u(\theta_i, q(\theta_i)) \geq \sum_{i=1}^{n} u(\theta_i, q(\theta_{i+1})).
$$

This is exactly the cyclical monotonicity condition in Rochet (1987).

When $X = \Theta$ is interpreted as the set of an agent's true types and $Y = \Theta$ interpreted as the set of reported types, the distribution $\pi^*$ constructed in the previous paragraph can be interpreted as the agent's truthful reporting strategy. Based on this interpretation, Rahman (2010) uses the duality between 1 and 3 to show that the incentive compatibility of truthful reporting subject to quota constraints is equivalent to implementability with transfers.

## 4.2 Finite-Sample Approximation

As discussed in Section 2.1, we interpret our model as one where the Sender designs a test that assigns scores to a large population of realized $\theta$'s; in particular, our model abstracts away from sampling variation, so there is no uncertainty in the population's realized type distribution. In this section we explicitly allow sampling variation by considering a finite-sample model where the Sender observes $N$ random i.i.d. draws from $\Theta$, and assigns each realized $\theta$ a message

$m \in M$, subject to certain message quotas—in particular, these message quotas substitute for the Sender's commitment to message distributions in the continuum model. We will show that credible and R-IC profiles in our continuum model are approximated by credible and R-IC profiles in the discrete model when sample size is large.

Consider a finite i.i.d sample of size $N$ drawn from the type space $\Theta$ according to the prior distribution $\mu_0$. The set of all possible empirical distributions over $\Theta$ in this $N$-sample captures the sampling variations in the realized type distribution, and can be written as

$$\mathcal{F}_\Theta^N = \left\{ f/N : f \in \mathbb{N}^{|\Theta|}, \sum_{\theta \in \Theta} f(\theta) = N \right\}.$$

The Sender assigns each realized $\theta$ in the $N$-sample a message $m \in M$, which leads to an $N$-sample of messages. Let

$$\mathcal{F}_M^N = \left\{ f/N : f \in \mathbb{N}^{|M|}, \sum_{m \in M} f(m) = N \right\}$$

denote the set of $N$-sample empirical distributions over messages. Lastly, for a pair of state and message distributions $(\mu, \nu)$, let

$$X^N(\mu, \nu) = \left\{ f/N : f \in \mathbb{N}^{|\Theta| \times |M|}, \sum_\theta f(\theta, \cdot) = N\nu(\cdot), \sum_m f(\cdot, m) = N\mu(\cdot) \right\}$$

denote the set of $N$-sample empirical joint distributions over states and messages that have marginals $\mu$ and $\nu$. Notice that $X^N(\mu, \nu) \neq \emptyset$ if and only if $\mu \in \mathcal{F}_\Theta^N$ and $\nu \in \mathcal{F}_M^N$.[14]

Let us now define the $N$-sample analogue of credible and R-IC profiles. We consider a Sender who assigns a message $m \in M$ to each realized $\theta \in \Theta$ subject to a message quota $\nu^N \in \mathcal{F}_M^N$. An $N$-sample profile is therefore a triple $(\nu^N, \phi^N, \sigma^N)$, where $\phi^N : \mathcal{F}_\Theta^N \to \Delta(\Theta \times M)$ is a Sender's strategy that takes every realized empirical distribution over states $\mu^N \in \mathcal{F}_\Theta^N$ to a joint distribution $\phi^N(\mu^N) \in X^N(\mu^N, \nu^N)$; meanwhile, $\sigma^N : M \to A$ is a Receiver's strategy that assigns an action to each observed message.[15]

The definitions of Sender credibility and Receiver incentive compatibility in the $N$-sample setting mirror those in our continuum model. In particular, we say an $N$-sample profile $(\nu^N, \phi^N, \sigma^N)$ is credible if for each realized empirical distribution over $\Theta$, the Sender always chooses an optimal assignment of messages subject to the message quotas specified in $\nu^N$:

---

[14]Notice that for any $f \in \mathbb{N}^{|\Theta| \times |M|}$, the sum of any row or column has to be integer, so $X^N(\mu, \nu) = \emptyset$ if either $\mu \notin \mathcal{F}_\Theta^N$ or $\nu \notin \mathcal{F}_M^N$. On the other hand if $\mu \in F_\Theta^N$ and $\nu \in F_M^N$, a $\lambda \in X^N(\mu, \nu)$ can be constructed from the so-called Northwest corner rule.

[15]Note that our formulation of the Sender's strategy assumes that the Sender conditions her strategy only on the empirical distribution of the realized $N$ samples, and ignores the identity of each individual sample point.

$(\nu^N, \phi^N, \sigma^N)$ is credible if for every $\mu^N \in \mathcal{F}_\Theta^N$,

$$\phi(\mu^N) \in \underset{\lambda^N \in X(\mu^N, \nu^N)}{\arg\max} \sum_{\theta, m} \lambda^N(\theta, m) u_S(\theta, \sigma^N(m)).$$

We say the $N$-sample profile $(\nu^N, \phi^N, \sigma^N)$ is Receiver incentive compatible (R-IC) if $\sigma^N$ best-responds to the Sender's strategy $\phi^N$. In particular, let $P^N$ denote the probability distribution over $\mathcal{F}_\Theta^N$ induced by i.i.d. draws from the prior distribution $\mu_0 \in \Delta(\Theta)$, and let $\phi^N(\theta, m|\mu^N)$ be the probability assigned to $(\theta, m)$ in the joint distribution $\phi^N(\mu^N)$ chosen by the Sender. The profile $(\nu^N, \phi^N, \sigma^N)$ is R-IC if

$$\sigma^N \in \underset{\sigma': M \to A}{\arg\max} \sum_{\mu^N \in \mathcal{F}_\Theta^N} P^N(\mu^N) \sum_{\theta, m} \phi(\theta, m|\mu^N) u_R(\theta, \sigma'(m)).$$

Proposition 5 below shows that credible and R-IC profiles in the continuum model are approximated by credible and R-IC profiles in the $N$-sample model, provided $N$ is sufficiently large. Note that in the second statement in Proposition 5, we distinguish a strictly credible profile $(\lambda^*, \sigma^*)$ in the continuum model as one where $\lambda^*$ is the unique maximizer in Definition 1; similarly, $(\lambda^*, \sigma^*)$ is strictly R-IC if $\sigma^*$ is the unique maximizer in Definition 2.

**Proposition 5.** *1. Let $(\lambda^*, \sigma^*)$ be a profile in the continuum model. If for every $\varepsilon > 0$, there exists a finite credible profile $(\nu^N, \phi^N, \sigma^N)$ for some sample size $N$, such that $|\nu^N - \lambda_M^*| < \varepsilon$, $|\sigma^N - \sigma^*| < \varepsilon$ and $P(|\phi^N(F_\Theta^N) - \lambda^*| < \varepsilon) > 1 - \varepsilon$, then $(\lambda^*, \sigma^*)$ is credible and R-IC.*

*2. Suppose $(\lambda^*, \sigma^*)$ is a strictly credible and strictly R-IC profile in the continuum model, then for each $\varepsilon > 0$ there exists a finite-sample credible and R-IC profile $(\nu^N, \phi^N, \sigma^N)$ such that $|\nu^N - \lambda_M^*| < \varepsilon$, $|\sigma^N - \sigma^*| < \varepsilon$ and $P(|\phi^N(F_\Theta^N) - \lambda^*| < \varepsilon) > 1 - \varepsilon$.*

The first statement in Proposition 5 is analogous to the upper-hemicontinuity of Nash equilibrium correspondences: if a profile $(\lambda^*, \sigma^*)$ in the continuous model can be arbitrarily approximated by credible and R-IC profiles in the finite model, then profile $(\lambda^*, \sigma^*)$ must itself be credible and R-IC. Conversely, the second statement in Proposition 5 can be interpreted in a way similar to the lower-hemicontinuity of strict Nash equiliria: if a profile $(\lambda^*, \sigma^*)$ in the continuous model is strictly credible and strictly R-IC, then it can be arbitrarily approximated by credible and R-IC profiles in the finite model.

## 4.3 Extensive-Form Foundation for Credibility

Our solution concept analyzes credible persuasion through the lens of a partial-commitment model, which is motivated by the observability of the Sender's message distribution. To formalize this connection, we propose an extensive-form game between the Sender and the Receiver where the Receiver observes the message distribution of each test. We show that the set of pure-strategy subgame perfect Nash equilibria of this extensive-form game corresponds to the set of profiles $(\lambda, \sigma)$ that are credible, R-IC, and give the Sender higher than her worst no-information payoff.

Consider the following game between the Sender and the Receiver:

1. The Sender chooses a test $\lambda \in \Delta(\Theta \times M)$ which satisfies $\lambda_\Theta = \mu_0$;
2. Nature draws a pair of state and message $(\theta, m)$ according to $\lambda$;
3. The Receiver observes $m$, as well as the distribution of messages induced by $\lambda$, $\lambda_M \in \Delta(M)$, then chooses an action $a \in A$.

The Sender's strategy set is $\Lambda \equiv \{\lambda \in \Delta(\Theta \times M) : \lambda_\Theta = \mu_0\}$, and the Receiver's strategy set is $\Xi = \{\rho : \Delta(M) \times M \to A\}$, where the first argument is the distribution of messages and the second argument is the message. The Sender's payoff is

$$U_S(\lambda, \rho) = \sum_\theta \sum_m \lambda(\theta, m) u_S\big(\theta, \rho(\lambda_M, m)\big)$$

while the Receiver's payoff is

$$U_R(\lambda, \rho) = \sum_\theta \sum_m \lambda(\theta, m) u_R\big(\theta, \rho(\lambda_M, m)\big)$$

The solution concept is pure-strategy subgame perfect Nash equilibrium (SPNE).[16] Notice that in the extensive-form game above, after the Sender chooses a degenerate test that always sends the a single message, the decision node for Nature forms the initial node of a proper subgame. In fact, these are also the only proper subgames in the extensive-form game, where subgame perfection has bite.

We call a profile $(\lambda, \sigma)$, as defined in Section 2.1, a *pure-strategy SPNE outcome* of the extensive-form game above if there exists a pure-strategy SPNE $(\lambda, \rho)$ of the extensive-form game such that $\rho(\lambda_M, m) = \sigma(m)$ for all $m \in M$. The following result relates the SPNE of this extensive-form game to our solution concept.

**Proposition 6.** *A profile $(\lambda, \sigma)$ is a pure-strategy SPNE outcome of the extensive-form game if and only if*

---

[16]Strengthening the solution concept to PBE does not change our result.

1. $(\lambda, \sigma)$ is credible and R-IC; that is, $(\lambda, \sigma)$ satisfies *Definitions 1* and *2*.

2. *The Sender's payoff from $(\lambda, \sigma)$ is greater than her lowest no-information payoff:*

$$\sum_\theta \sum_m \lambda(\theta, m) u_S(\theta, \sigma(m)) \geq \min_{a \in A_0} \sum_\theta \mu_0(\theta) u_S(\theta, a),$$

where $A_0 = \arg\max_{a \in A} \sum_\theta \mu_0(\theta) u_R(\theta, a)$ is the Receiver's best-response set to the prior belief $\mu_0$.

Proposition 6 shows that if one focuses on profiles where the Sender is weakly better off than disclosing no information (as we do in this paper), the extensive-form game rationalizes our definition for credibility. The reason that the Sender's payoff must be higher than her no-information payoff in the extensive-form game is that in any equilibrium, if she deviates to a no-information test $\lambda = \mu_0 \times \delta_{m_0}$, the ensuing decision node forms the initial node of a proper subgame. Subgame perfection then demands that the Receiver plays a best response to his prior, which in turn ensures that the Sender obtains a no-information payoff following this deviation. Therefore, the Sender's equilibrium payoff must be weakly higher than her worst no-information payoff.

# 5  Conclusion

This paper offers a new notion of credibility for persuasion problems. We model a Sender who can commit to a test only up to the details that are observable to the Receiver. The Receiver does not observe the chosen test but observes the distribution of messages. This leads to a model of partial commitment where the Sender can undetectably deviate to tests that induce the same distribution of messages. Our framework characterizes when, given the Receiver's best response, the Sender has no profitable deviation.

We show that this consideration eliminates the prospects for credible persuasion in settings with adverse selection. In other settings, we show that the requirement is compatible with the Sender still benefiting from persuasion. More generally, we show that our requirement translates to a cyclical monotonicity condition on the induced distribution of states and actions.

In addition to the theoretical findings above, our work has several applied implications. The first is that the commitment assumption commonly made in the Bayesian persuasion literature may be innocuous in some applications. For example, Xiang (2021) uses a Bayesian persuasion framework to empirically study information transmission in the physician-patient relationship, where the physician is assumed to commit to a recommendation policy that is

1. $(\lambda, \sigma)$ is credible and R-IC; that is, $(\lambda, \sigma)$ satisfies *Definitions 1* and *2*.

2. *The Sender's payoff from $(\lambda, \sigma)$ is greater than her lowest no-information payoff:*

$$\sum_\theta \sum_m \lambda(\theta, m) u_S(\theta, \sigma(m)) \geq \min_{a \in A_0} \sum_\theta \mu_0(\theta) u_S(\theta, a),$$

where $A_0 = \arg\max_{a \in A} \sum_\theta \mu_0(\theta) u_R(\theta, a)$ is the Receiver's best-response set to the prior belief $\mu_0$.

Proposition 6 shows that if one focuses on profiles where the Sender is weakly better off than disclosing no information (as we do in this paper), the extensive-form game rationalizes our definition for credibility. The reason that the Sender's payoff must be higher than her no-information payoff in the extensive-form game is that in any equilibrium, if she deviates to a no-information test $\lambda = \mu_0 \times \delta_{m_0}$, the ensuing decision node forms the initial node of a proper subgame. Subgame perfection then demands that the Receiver plays a best response to his prior, which in turn ensures that the Sender obtains a no-information payoff following this deviation. Therefore, the Sender's equilibrium payoff must be weakly higher than her worst no-information payoff.

# 5  Conclusion

This paper offers a new notion of credibility for persuasion problems. We model a Sender who can commit to a test only up to the details that are observable to the Receiver. The Receiver does not observe the chosen test but observes the distribution of messages. This leads to a model of partial commitment where the Sender can undetectably deviate to tests that induce the same distribution of messages. Our framework characterizes when, given the Receiver's best response, the Sender has no profitable deviation.

We show that this consideration eliminates the prospects for credible persuasion in settings with adverse selection. In other settings, we show that the requirement is compatible with the Sender still benefiting from persuasion. More generally, we show that our requirement translates to a cyclical monotonicity condition on the induced distribution of states and actions.

In addition to the theoretical findings above, our work has several applied implications. The first is that the commitment assumption commonly made in the Bayesian persuasion literature may be innocuous in some applications. For example, Xiang (2021) uses a Bayesian persuasion framework to empirically study information transmission in the physician-patient relationship, where the physician is assumed to commit to a recommendation policy that is

observable to patients. But in practice, patients may observe the distribution of recommendations and not the recommendation policy. Our results imply that in her context, this is enough: knowing the distribution of recommendations suffices because both the physician and the patients' payoffs are supermodular, and the patients face a binary decision, so the optimal full-commitment policy is credible according to our Proposition 3.

Our work also speaks to why certain industries (such as education) can effectively disclose information by utilizing their own rating systems, while some other industries (such as car dealerships) must resort to other means to addressing asymmetric information, such as third-party certification or warranties. Our results provide a rationale: in industries that exhibit adverse selection, the informed party cannot credibly disclose information through its own ratings even if it wishes to do so.

The notion of credibility we consider in this paper is motivated by the observability of the Sender's message distribution. In some settings, the Receiver may observe more than the distribution of messages; for example, she may observe some further details of the test, such as how some states of the world map into messages. In other settings, the Receiver may observe less; e.g., she may see the average grade, but not its distribution. To capture these various cases, one would then formulate the problem of "detectable" deviations differently. We view it to be an interesting direction for future research to understand how different notions of detectability map into different conditions on the outcome distribution.

# References

Akbarpour, Mohammad and Shengwu Li. 2020. "Credible Auctions: A Trilemma." *Econometrica* 88 (2):425–467.

Akerlof, George A. 1970. "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism." *Quarterly Journal of Economics* 84 (3):488–500.

Alonso, Ricardo and Odilon Câmara. 2016. "Persuading Voters." *American Economic Review* 106 (11):3590–3605.

Beiglböck, Mathias, Martin Goldstern, Gabriel Maresch, and Walter Schachermayer. 2009. "Optimal and Better Transport Plans." *Journal of Functional Analysis* 256 (6):1907–1927.

Bergemann, Dirk and Stephen Morris. 2016. "Bayes Correlated Equilibrium and the Comparison of Information Structures in Games." *Theoretical Economics* 11 (2):487–522.

Best, James and Daniel Quigley. 2020. "Persuasion for the Long Run." Working Paper.

Brocas, Isabelle and Juan D Carrillo. 2007. "Influence Through Ignorance." *The RAND Journal of Economics* 38 (4):931–947.

Brualdi, Richard A. 2006. *Combinatorial Matrix Classes*, vol. 13. Cambridge University Press.

Chakraborty, Archishman and Rick Harbaugh. 2007. "Comparative Cheap Talk." *Journal of Economic Theory* 132 (1):70–94.

———. 2010. "Persuasion by Cheap Talk." *American Economic Review* 100 (5):2361–82.

Crawford, Vincent P. and Joel Sobel. 1982. "Strategic Information Transmission." *Econometrica* 50 (6):1431–1451.

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

Escobar, Juan F and Juuso Toikka. 2013. "Efficiency in Games with Markovian Private Information." *Econometrica* 81 (5):1887–1934.

Frankel, Alexander. 2014. "Aligned Delegation." *American Economic Review* 104 (1):66–83.

Gitmez, A. Arda and Pooya Molavi. 2022. "Polarization and Media Bias." *arXiv preprint arXiv:2203.12698* .

Goldstein, Itay and Yaron Leitner. 2018. "Stress Tests and Information Disclosure." *Journal of Economic Theory* 177:34–69.

Ivanov, Maxim. 2020. "Optimal Monotone Signals in Bayesian Persuasion Mechanisms." *Economic Theory* 72:955–1000.

Jackson, Matthew O. and Hugo F. Sonnenschein. 2007. "Overcoming Incentive Constraints by Linking Decisions." *Econometrica* 75 (1):241–257.

Kamenica, Emir. 2019. "Bayesian Persuasion and Information Design." *Annual Review of Economics* 11:249–272.

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

Kartik, Navin and Weijie Zhong. 2019. "Lemonade from Lemons: Information Design and Adverse Selection." Working Paper.

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

Kolotilin, Anton and Hongyi Li. 2020. "Relational Communication." *Theoretical Economics* (Forthcoming).

Kuvalekar, Aditya, Elliot Lipnowski, and Joao Ramos. 2021. "Goodwill in Communication." *Journal of Economic Theory* (Forthcoming).

Lipnowski, Elliot and Doron Ravid. 2020. "Cheap Talk with Transparent Motives." *Econometrica* 88 (4):1631–1660.

Lipnowski, Elliot, Doron Ravid, and Denis Shishkin. 2021. "Persuasion via Weak Institutions." Working Paper.

Margaria, Chiara and Alex Smolin. 2018. "Dynamic Communication with Biased Senders." *Games and Economic Behavior* 110:330–339.

Mas-Colell, Andreu, Michael Dennis Whinston, and Jerry R. Green. 1995. *Microeconomic Theory*, vol. 1. Oxford University Press New York.

Mathevet, Laurent, David Pearce, and Ennio Stacchetti. 2022. "Reputation for A Degree of Honesty." Working Paper.

Matsushima, Hitoshi, Koichi Miyazaki, and Nobuyuki Yagi. 2010. "Role of Linking Mechanisms in Multitask Agency with Hidden Information." *Journal of Economic Theory* 145 (6):2241–2259.

Meng, Delong. 2021. "On the Value of Repetition for Communication Games." *Games and Economic Behavior* 127:227–246.

Mensch, Jeffrey. 2021. "Monotone Persuasion." *Games and Economic Behavior* (Forthcoming).

Min, Daehong. 2021. "Bayesian Persuasion under Partial Commitment." *Economic Theory* 72:743–764.

Nguyen, Anh and Teck Yong Tan. 2021. "Bayesian Persuasion with Costly Messages." *Journal of Economic Theory* 193:105212.

Ostrovsky, Michael and Michael Schwarz. 2010. "Information Disclosure and Unraveling in Matching Markets." *American Economic Journal: Microeconomics* 2 (2):34–63.

Pei, Harry. 2020. "Repeated Communication with Private Lying Cost." Working Paper.

Perez-Richet, Eduardo and Vasiliki Skreta. 2021. "Test Design under Falsification." Working Paper.

Rahman, David. 2010. "Detecting Profitable Deviations." Working Paper.

Rayo, Luis and Ilya Segal. 2010. "Optimal Information Disclosure." *Journal of Political Economy* 118 (5):949–987.

Renault, Jérôme, Eilon Solan, and Nicolas Vieille. 2013. "Dynamic Sender–Receiver games." *Journal of Economic Theory* 148 (2):502–534.

Rochet, Jean-Charles. 1987. "A Necessary and Sufficient Condition for Rationalizability in a Quasi-linear Context." *Journal of Mathematical Economics* 16 (2):191–200.

Santambrogio, Filippo. 2015. "Optimal Transport for Applied Mathematicians." *Birkäuser, NY* 55 (58-63):94.

Taneva, Ina. 2019. "Information Design." *American Economic Journal: Microeconomics* 11 (4):151–85.

Topkis, Donald M. 2011. *Supermodularity and Complementarity*. Princeton University Press.

Villani, Cédric. 2008. *Optimal Transport: Old and New*, vol. 338. Springer Science & Business Media.

Xiang, Jia. 2021. "Physicians as Persuaders: Evidence from Hospitals in China." Working Paper.

# A Appendix

## A.1 Proof of Theorem 1

The following lemma, which will play a key role in the proof of Theorem 1, is a finite version of Theorem 5.10 of Villani (2008).

**Lemma 2.** *Suppose both $X$ and $Y$ are finite sets, and $u : X \times Y \to \mathbb{R}$ is a real function. Let $p \in \Delta(X)$ and $q \in \Delta(Y)$ be two probability measure on $X$ and $Y$ respectively, and $\Pi(p, q)$ be the set of joint probability measure on $X \times Y$ such that the marginals on $X$ and $Y$ are $p$ and $q$. The following two statements are equivalent:*

*1. $\pi^* \in \arg\max_{\pi \in \Pi(p,q)} \sum_{x,y} \pi(x, y) u(x, y)$;*

*2. $\pi^*$ is u-cyclically monotone. That is, for any $n$ and $(x_1, y_1), ..., (x_n, y_n) \in supp(\pi^*)$,*

$$\sum_{i=1}^{n} u(x_i, y_i) \geq \sum_{i=1}^{n} u(x_i, y_{i+1})$$

*where $y_{n+1} \equiv y_1$.*

*Proof of Theorem 1.* For the "if" direction, suppose $\pi$ is $u_R$-obedient, $u_S$-cyclically monotone, and satisfies $\pi_\Theta = \mu_0$. The proof is by construction.

Since $\pi_\Theta = \mu_0$, we can construct a test $(M, \lambda^*)$ by setting $M = A$ and $\lambda^* = \pi$; furthermore, let $\sigma^*$ be the identity map from $M$ to $A$. It is straightforward to see that the profile $(\lambda^*, \sigma^*)$ induces the outcome distribution $\pi$. We show that $(\lambda^*, \sigma^*)$ is credible. First, since $\pi$ is $u_R$-obedient, we have that for each $a \in A$,

$$a \in \arg\max_{a'} \sum_{\Theta} u_R(\theta, a') \, \pi(\theta, a).$$

Since $\sigma^*$ is an identity map, it follows that for each $m \in M$,

$$\sigma^*(m) \in \arg\max_{a'} \sum_{\Theta} u_R(\theta, a') \, \pi(\theta, \sigma^*(m)).$$

Furthermore, since $\lambda^* = \pi$ and $\sigma^*$ is injective, we have $\lambda^*(\theta, m) = \pi(\theta, \sigma^*(m))$ for all $\theta \in \Theta$ and $m \in M$. So

$$\sigma^* \in \arg\max_{\sigma: M \to A} \sum_{\Theta \times M} u_R(\theta, \sigma(m)) \, \lambda^*(\theta, m),$$

which means $\sigma^*$ is a best response to $\lambda^*$.

It remains to show that the Sender does not benefit from choosing any other test in $\Lambda(\mu_0, \lambda_M^*)$. Observe that since $\pi$ is $u_S$−cyclically monotone, every sequence $(\theta_1, a_1), \ldots, (\theta_n, a_n)$

29

in supp($\pi$) where $a_{n+1} \equiv a_1$ satisfies

$$\sum_{i=1}^{n} u_S(\theta_i, a_i) \geq \sum_{i=1}^{n} u_S(\theta_i, a_{i+1}).$$

Since $\lambda^* = \pi$ and $\sigma^*$ is the identity mapping, this further implies

$$\sum_{i=1}^{n} u_S(\theta_i, \sigma^*(m_i)) \geq \sum_{i=1}^{n} u_S(\theta_i, \sigma^*(m_{i+1}));$$

for every sequence $(\theta_1, m_1), \ldots, (\theta_n, m_n) \in$ supp($\lambda^*$) with $m_{n+1} = m_1$. In addition, $\lambda_\theta^* = \mu_0$ and $\lambda_M^* = \lambda_M^*$ by construction. By Lemma 2, $\lambda^*$ satisfies

$$\lambda^* \in \underset{\lambda \in \Lambda(\mu_0, \lambda_M^*)}{\arg\max} \sum_{\Theta \times M} u_S(\theta, \sigma(m)) \, \lambda(\theta, m)$$

which means $\lambda^*$ is Sender optimal conditional on its message distribution.

For the "only if" direction, suppose $\pi$ is stable and thus induced by a credible and R-IC profile $(\lambda^*, \sigma^*)$. Since $\sigma^*$ best responds to the messages from $\lambda^*$, the $u_R$-obedience of $\pi$ follows from Bergemann and Morris (2016). It remains to show that $\pi$ is $u_S$-cyclical monotone. Suppose by contradiction that $\pi$ is not $u_S$-cyclically monotone, which implies that there exists a sequence $(\theta_1, a_1), \ldots, (\theta_n, a_n) \in$ supp($\pi$) such that

$$\sum_{i=1}^{n} u_S(\theta_i, a_i) < \sum_{i=1}^{n} u_S(\theta_i, a_{i+1}),$$

where $a_{n+1} = a_1$. Since $\pi$ is induced by $(\lambda^*, \sigma^*)$, for each $i = 1, \ldots, n$ there exists $m_i$ such that $m_i \in \sigma^{*-1}(a_i)$ and $(\theta_i, m_i) \in$ supp($\lambda^*$), so we have a sequence $(\theta_1, m_1), \ldots, (\theta_n, m_n) \in$ supp($\lambda^*$) that satisfies

$$\sum_{i=1}^{n} u_S(\theta_i, \sigma^*(m_i)) < \sum_{i=1}^{n} u_S(\theta_i, \sigma^*(m_{i+1})), \tag{3}$$

where $m_{n+1} = m_1$. Define $v(\theta, m) \equiv u_S(\theta, \sigma^*(m))$. Since $(\lambda^*, \sigma^*)$ is credible, we have

$$\lambda^* \in \underset{\lambda \in \Lambda(\mu_0, \lambda_M^*)}{\arg\max} \sum_{\Theta \times M} v(\theta, m) \lambda(\theta, m).$$

Lemma 2 implies that $\lambda^*$ is $v$-cyclically monotone. Since $(\theta_1, m_1), \ldots, (\theta_n, m_n)$ is in supp($\lambda^*$),

the $v$-cyclical monotonicity of $\lambda^*$ implies

$$\sum_{i=1}^{n} u_S(\theta_i, \sigma^*(m_i)) \geq \sum_{i=1}^{n} u_S(\theta_i, \sigma^*(m_{i+1}))$$

where $m_{n+1} = m_1$, which is a contradiction to (3). So $\pi$ must be $u_S$-cyclically monotone.

$\square$

## A.2 Proof of Lemma 1

**Lemma 3.** *Let $t : \{1, \ldots, n\} \to \{1, \ldots, n\}$ be a bijection. Suppose $t$ is not the identity mapping, then there exists $k$ such that $t(k) > k$ and $t(t(k)) < t(k)$.*

*Proof.* Suppose by contradiction that for every $k$ such that $t(k) > k$, $t(t(k)) \geq t(k)$. Notice that since $t$ is a bijection, $t(t(k)) \neq t(k)$ (otherwise $t(k) = k$ contradicting $t(k) > k$), so for every $k$ such that $t(k) > k$, $t(t(k)) \geq t(k) + 1$.

Since $t$ is not the identity mapping, there exists $k_1$ such that $t(k_1) > k_1$ or equivalently $t(k_1) \geq k_1 + 1$. Define iteratively that $k_j = t(k_{j-1})$ for $j = 2, \ldots, n$, we have $k_j - k_{j-1} \geq 1$. Then we have $k_n \geq k_1 + n > n$, which is contradiction. So there exists $k$ such that $t(k) > k$ and $t(t(k)) < t(k)$. $\square$

First, we show that comonotonicity implies $u_S$-cyclical monotonicity when $u_S$ is supermodular. Suppose an outcome distribution $\pi \in \Delta(\Theta \times A)$ is comonotone, then $\text{supp}(\pi)$ is totally ordered. Take any sequence $(\theta_1, a_1), \ldots, (\theta_n, a_n) \in \text{supp}(\pi)$ and assume without loss of generality that $(\theta_i, a_i)$ is increasing in $i \in \{1, \ldots, n\}$. We will show that for any permutation $t : \{1, \ldots, n\} \to \{1, \ldots, n\}$,

$$u_S(\theta_1, a_1) + \ldots + u_S(\theta_n, a_n) \geq u_S(\theta_1, a_{t(1)}) + \ldots + u_S(\theta_n, a_{t(n)}),$$

which then proves the statement. In particular, for each permutation $t$, let $v(t) \equiv u_S(\theta_1, a_{t(1)}) + \ldots + u_S(\theta_n, a_{t(n)})$ denote the value obtained from summing $u_S$ according to the state-action pairings in $t$ and let $I$ denote the identity map. We show that $v(I) \geq v(t)$ for every permutation $t$.

To this end, take any permutation $t$ that is not an identity mapping, and let $l(t)$ denote the number of fixed points of $t$ (which may be zero). By Lemma 3, there exists $k^*$ such that $t(k^*) > k^*$ and $t(t(k^*)) < t(k^*)$. The supermodularity of $u_S$ implies

$$u_S(\theta_{t(k^*)}, a_{t(k^*)}) + u_S(\theta_{k^*}, a_{t(t(k^*))}) \geq u_S(\theta_k, a_{t(k^*)}) + u_S(\theta_{t(k^*)}, a_{t(t(k^*))}). \tag{4}$$

31

Define a new permutation $\hat{t}$ so that $k$ is mapped to $t(t(k))$ while $t(k)$ is mapped to $t(k)$, while all other pairings remain unchanged. Formally,

$$\hat{t}(k) = \begin{cases} t(k) & \text{for all } k \neq k^*, t(k^*) \\ t(t(k^*)) & \text{if } k = k^* \\ t(k^*) & \text{if } k = t(k^*) \end{cases}$$

By (4), we have

$$u_S(\theta_1, a_{\hat{t}(1)}) + \ldots + u_S(\theta_n, a_{\hat{t}(n)}) \geq u_S(\theta_1, a_{t(1)}) + \ldots + u_S(\theta_n, a_{t(n)}),$$

so we have constructed another permutation $\hat{t}$ with $v(\hat{t}) \geq v(t)$ and $l(\hat{t}) = l(t) + 1$. Each time we iterate the process above, $v(.)$ weakly increases while the number of fixed points increases by one. Since $n < \infty$, the iteration terminates at the identity map $I$, so $v(I) \geq v(t)$ for every permutation $t$.

Next, suppose $u_S$ is strictly supermodular. We will show that $u_S$-cyclical monotonicity implies comonotonicity. Towards a contradiction, suppose that an outcome distribution $\pi$ is $u_S$-cyclically monotone but not comonotone. Then there exists $(\theta, a), (\theta', a') \in \text{supp}(\pi)$ such that $\theta < \theta'$, $a > a'$. Since $u_S$ is strictly supermodular,

$$u_S(\theta, a) + u_S(\theta', a') < u_S(\theta, a') + u_S(\theta', a)$$

which is a contradiction to the $u_S$-cyclically monotonicity of $\pi$ when $(\theta_1, a_1) = (\theta, a)$ and $(\theta_2, a_2) = (\theta', a')$.

## A.3   Proof of Proposition 1

Let $\pi$ be a stable outcome distribution, and suppose by contradiction that there exists two distinct actions $a_1, a_2 \in \text{supp}(\pi_a)$, say $a_1 < a_2$. Let $I_1 \equiv \{\theta \in \Theta | \pi(\theta, a_1) > 0\}$ and $I_2 \equiv \{\theta \in \Theta | \pi(\theta, a_2) > 0\}$ be the states associated with $a_1$ and $a_2$ in the support of $\pi$, respectively. By Theorem 1, since $\pi$ is stable, it must be $u_R$-obedient, which implies

$$\sum_{\theta \in I_1} [u_R(\theta, a_1) - u_R(\theta, a_2)] \frac{\pi(\theta, a_1)}{\pi_a(a_1)} \geq 0 \geq \sum_{\theta' \in I_2} [u_R(\theta', a_1) - u_R(\theta', a_2)] \frac{\pi(\theta', a_2)}{\pi_a(a_2)} \quad (5)$$

Furthermore, since $u_S$ is strictly supermodular, $\pi$ is also comonotone by Theorem 1 and Lemma 1, so any $\theta \in I_1$ and $\theta' \in I_2$ satisfies $\theta \leq \theta'$. Since $u_R$ is submodular, we have

$u_R(\theta, a_1) - u_R(\theta, a_2) \leq u_R(\theta', a_1) - u_R(\theta', a_2)$ for all $\theta \in I_1$ and $\theta' \in I_2$, which implies

$$\max_{\theta \in I_1} \left\{ u_R(\theta, a_1) - u_R(\theta, a_2) \right\} \leq \min_{\theta' \in I_2} \left\{ u_R(\theta', a_1) - u_R(\theta', a_2) \right\}.$$

So

$$
\begin{aligned}
\sum_{\theta \in I_1} [u_R(\theta, a_1) - u_R(\theta, a_2)] \frac{\pi(\theta, a_1)}{\pi_a(a_1)} &\leq \max_{\theta \in I_1} \left\{ u_R(\theta, a_1) - u_R(\theta, a_2) \right\} \\
&\leq \min_{\theta' \in I_2} \left\{ u_R(\theta', a_1) - u_R(\theta', a_2) \right\} \qquad (6) \\
&\leq \sum_{\theta' \in I_2} [u_R(\theta', a_1) - u_R(\theta', a_2)] \frac{\pi(\theta', a_2)}{\pi_a(a_2)}
\end{aligned}
$$

Combining (5) and (6), we have

$$\sum_{\theta \in I_1} [u_R(\theta, a_1) - u_R(\theta, a_2)] \frac{\pi(\theta, a_1)}{\pi_a(a_1)} = \max_{\theta \in I_1} \left\{ u_R(\theta, a_1) - u_R(\theta, a_2) \right\} = 0$$

and

$$\sum_{\theta' \in I_2} [u_R(\theta', a_1) - u_R(\theta', a_2)] \frac{\pi(\theta', a_2)}{\pi_a(a_2)} = \min_{\theta' \in I_2} \left\{ u_R(\theta', a_1) - u_R(\theta', a_2) \right\} = 0$$

So $u_R(\theta, a_1) = u_R(\theta, a_2)$ for all $\theta \in I_1 \cup I_2$.

Since the argument above works for any $a_1, a_2 \in \text{supp}(\pi_a)$, it implies $u_R(\theta, a) = u_R(\theta, a')$ for all $\theta \in \Theta$ and all $a, a' \in \text{supp}(\pi_a)$. However, this is a contradiction since by assumption, there exists no $a, a' \in A$ such that $a \neq a'$ and $u_R(\theta, a) = u_R(\theta, a')$ for all $\theta$.

Therefore $\text{supp}(\pi_a)$ must be a singleton, denoted by $a^*$. Then $u_R$-obeidence implies $a^* \in \arg\max_{a \in A} \sum_\theta \mu_0(\theta) u(\theta, a)$. So $\pi$ is a no-information outcome.

## A.4   Proof of Proposition 2

*Proof of statement 1.* For each $a \in A$, let

$$P_a \equiv \{ \mu \in \Delta(\Theta) | \sum_\theta \mu(\theta) u_R(\theta, a) > \sum_\theta \mu(\theta) u_R(\theta, a'), \forall a' \neq a \}$$

which denotes the set of beliefs such that $a$ is the Receiver's strict best response. We prove our claim under the assumption that there exists $a^\circ \in A$ such that $\mu_0 \in P_{a^\circ}$ (i.e. $a^\circ$ is the unique best response to $\mu_0$). Later we will show that this assumption holds for generic priors.

When the Sender's test is uninformative, the Receiver best responds to the Sender's mes-

sages by choosing $a^\circ$. The Sender's payoff is

$$v_0 \equiv \sum_{\theta \in \Theta} \mu_0(\theta) u_S(\theta, a^0).$$

We will show that there exists a stable outcome distribution that gives the Sender a higher payoff than $v_0$.

We consider the case where the sender benefits from persuasion, so $a^\circ \neq \bar{a}$, otherwise the Receiver is already choosing the sender's favourite action under the prior. For $\varepsilon$ sufficiently small, consider the outcome distribution $\pi^\varepsilon \in \Delta(\Theta \times A)$ defined by

$$\pi^\varepsilon(\theta, a) = \begin{cases} \mu_0(\theta) & \text{if } \theta \neq \bar{\theta}, a = a^\circ \\ \mu_0(\bar{\theta}) - \varepsilon & \text{if } (\theta, a) = (\bar{\theta}, a^\circ) \\ \varepsilon & \text{if } (\theta, a) = (\bar{\theta}, \bar{a}) \\ 0 & \text{otherwise .} \end{cases}$$

We will show that for $\varepsilon$ sufficiently small, $\pi^\varepsilon$ is stable and gives the Sender higher payoff than $v_0$.

It can be easily seen that the support of $\pi^\varepsilon$ is comonotone. Since $u_S$ is supermodular, $\pi^\varepsilon$ is $u_S$-cyclically monotone by Lemma 1.

Next we verify that for $\varepsilon$ sufficiently small, $\pi^\varepsilon$ satisfies $u_R$-obedience at the two actions $\{\bar{a}, a^\circ\}$. For $a^\circ$, note that since $\mu_0 \in P_{a^\circ}$, we have

$$\sum_{\theta \in \Theta} \mu_0(\theta) u(\theta, a^\circ) > \sum_{\theta \in \Theta} \mu_0(\theta) \pi(\theta, a') \text{ for all } a' \in A,$$

so for $\varepsilon$ sufficiently small,

$$\sum_{\theta \in \Theta} \mu_0(\theta) u(\theta, a^\circ) - \varepsilon u(\bar{\theta}, a^\circ) \geq \sum_{\theta \in \Theta} \mu_0(\theta) \pi(\theta, a') - \varepsilon u(\bar{\theta}, a') \text{ for all } a' \in A.$$

which means $\pi^\varepsilon$ satisfies $u_R$-obedience at $a^\circ$.

For $\bar{a}$, note that since every Receiver action is a best response to some belief (recall that this was assumed without loss of generality as explained in Section 2.4), there exists $\bar{\mu} \in \Delta(\Theta)$ such that $\bar{a} \in \arg\max_a \sum_\theta \bar{\mu}(\theta) u_R(\theta, a)$. So for every $a' \neq \bar{a}$,

$$\sum_\theta \bar{\mu}(\theta)[u_R(\theta, \bar{a}) - u_R(\theta, a')] \geq 0$$

Since $u_R$ is supermodular, $u_R(\theta, \bar{a}) - u_R(\theta, a')$ is increasing in $\theta$, so if a belief $\mu'$ first order

stochastically dominates $\bar{\mu}$, then

$$\sum_\theta \mu'(\theta)[u_R(\theta, \bar{a}) - u(\theta, a')] \geq \sum_\theta \bar{\mu}(\theta)[u_R(\theta, \bar{a}) - u(\theta, a')] \geq 0 \text{ for all } a' \neq \bar{a}.$$

In particular, the Dirac measure $\delta_{\bar{\theta}}$ first order stochastically dominates $\bar{\mu}$, so the inequality above implies

$$u_R(\bar{\theta}, \bar{a}) - u_R(\bar{\theta}, a') \geq 0 \text{ for all } a' \neq \bar{a}.$$

So $\bar{a} \in \arg\max_a u_R(\bar{\theta}, a)$, and $\pi^\varepsilon$ is $u_R$-obedient at action $\bar{a}$.

Finally, we show that the Sender obtains higher payoff from $\pi^\varepsilon$ than $v_0$. Note that since by our assumption, $u_S(\bar{\theta}, a') < u_S(\bar{\theta}, \bar{a})$ for all $a' \neq \bar{a}$, we have

$$\sum_{\theta, a} \pi^\varepsilon(\theta, a) u_S(\theta, a) = \sum_{\theta \neq \bar{\theta}} \mu_0(\theta) u_S(\theta, a^\circ) + (\mu_0(\bar{\theta}) - \varepsilon) u_S(\bar{\theta}, a^\circ) + \varepsilon u_S(\bar{\theta}, \bar{a})$$

$$> \sum_{\theta \neq \bar{\theta}} \mu_0(\theta) u_S(\theta, a^\circ) + (\mu_0(\bar{\theta}) - \varepsilon) u_S(\bar{\theta}, a^\circ) + \varepsilon u_S(\bar{\theta}, a^\circ)$$

$$= \sum_\theta \mu_0(\theta) u_S(\theta, a^\circ) = v_0.$$

Therefore, Sender receives a strictly higher payoff from $\pi^\varepsilon$ than $v_0$. This completes the proof.

The rest of the proof shows that the set $\Delta(\Theta)/\{\cup_{a \in A} P_a\}$ is negligible in $\Delta(\Theta)$.

Define $H_{a,a'} \equiv \{\mu \in \Delta(\Theta) | \sum_\theta \mu(\theta)(u_R(\theta, a) - u_R(\theta, a')) = 0\}$ for any $a \neq a'$. Since by assumption, $u_R(\cdot, a) - u_R(\cdot, a') \neq \mathbf{0}$, which implies $J_{a,a'} \equiv \{\mu \in \mathbb{R}^{|\Theta|} | \sum_\theta \mu(\theta)(u_R(\theta, a) - u_R(\theta, a')) = 0\}$ is a hyperplane in $\mathbb{R}^{|\Theta|}$. Notice that $H_{a,a'} = J_{a,a'} \cap \Delta(\Theta)$, which is the intersection of a hyperplane with a simplex. Since the hyperplane includes $\mathbf{0}$ and $\Delta(\Theta)$ doesn't, they have to either be parallel with no intersection, or their intersection is in a lower dimensional space, which has measure 0 in $\Delta(\Theta)$.

For any $\mu \in \Delta(\Theta)/\{\cup_{a \in A} P_a\}$, since the maximizer of $\sum_\theta \mu(\theta) u_R(\theta, a)$ is not unique, there exists $a, a'$ such that $\sum_\theta \mu(\theta)(u_R(\theta, a) - u_R(\theta, a')) = 0$. So $\Delta(\Theta)/\{\cup_{a \in A} P_a\} \subset \cup_{a \neq a'} H_{a,a'}$, which implies $\Delta(\Theta)/\{\cup_{a \in A} P_a\}$ is a negligible set in $\Delta(\Theta)$.

$\square$

*Proof of statement 2.* For any generic prior $\mu^\circ \in \cup_{a \in A} P_a$, either $\mu^\circ \notin P_{\underline{a}}$ or $\mu^\circ \notin P_{\bar{a}}$. We consider the case $\mu^\circ \notin P_{\bar{a}}$, and the other case can be shown symmetrically. Similar as the

previous argument, for $\varepsilon$ sufficiently small, consider the outcome distribution $\pi^{\varepsilon} \in \Delta(\Theta \times A)$:

$$\pi^{\varepsilon}(\theta, a) = \begin{cases} \mu_0(\theta) & \text{if } \theta \neq \bar{\theta}, a = a^{\circ} \\ \mu_0(\bar{\theta}) - \varepsilon & \text{if } (\theta, a) = (\bar{\theta}, a^{\circ}) \\ \varepsilon & \text{if } (\theta, a) = (\bar{\theta}, \bar{a}) \\ 0 & \text{otherwise} \end{cases}$$

As we have shown in the proof of statement 1, for $\varepsilon$ sufficiently small, $\pi^{\varepsilon}$ is stable, and gives the Sender higher payoff than $v_0$. Therefore, the sender benefits from credible persuasion.

$\square$

*Proof of statement 3.* Consider any fully revealing outcome distribution $\pi^* \in \Delta(\Theta \times A)$ which gives a strictly higher payoff to the Sender than every no-information outcome. Let $A^*(\theta) \equiv \arg\max_{a \in A} u_R(\theta, a)$ denote the Receiver's best response correspondence. By definition, for every $(\theta, a) \in \text{supp}(\pi^*)$, $a \in A^*(\theta)$. If $\pi^*$ is comonotone, then then from Theorem 1 and Lemma 1, $\pi^*$ is credible so the result follows. If $\pi^*$ is not comonotone, then there exists $(\theta, a)$, $(\theta', a')$ in the support of $\pi^*$ where $\theta > \theta'$ and $a < a'$. Let $d = \max\{\theta' - \theta | (\theta, a), (\theta', a') \in \text{supp}(\pi^*), \theta > \theta', a < a'\}$ denote the largest distance of states between those "non-monotone" pairs. Suppose $(\theta_1, a_1)$, $(\theta_2, a_2)$ is a pair that induces the largest distance, where $\theta_1 < \theta_2$ and $a_1 > a_2$.

Let $\varepsilon = \min\{\pi^*(\theta_1, a_1), \pi^*(\theta_2, a_2)\}$, and construct the following outcome distribution $\pi'$:

- $\pi'(\theta_1, a_1) = \pi^*(\theta_1, a_1) - \varepsilon$, $\pi'(\theta_2, a_2) = \pi^*(\theta_2, a_2) - \varepsilon$
- $\pi'(\theta_1, a_2) = \pi^*(\theta_1, a_2) + \varepsilon$, $\pi'(\theta_2, a_1) = \pi^*(\theta_2, a_1) + \varepsilon$
- $\pi'(\theta, a) = \pi^*(\theta, a)$ for any other $(\theta, a)$

For any $a \notin \{a_1, a_2\}$, the obedient constraint under $\pi'$ is the same as under $\pi^*$, so the obedient constraint still holds. For $a \in \{a_1, a_2\}$, we show that the obedient constraint is still satisfied.

Since $u_R(\theta, a)$ is supermodular, by Lemma 2.8.1 of Topkis (2011), $A^*(\theta)$ is increasing in $\theta$ in the induced set order. That is, for any $\theta > \theta'$, $a \in A^*(\theta)$, and $a' \in A^*(\theta')$, we have $\max\{a, a'\} \in A^*(\theta)$ and $\min\{a, a'\} \in A^*(\theta')$. Since $a_1 \in A^*(\theta_1)$ and $a_2 \in A^*(\theta_2)$, we have $a_1 \in A^*(\theta_2)$ and $a_2 \in A^*(\theta_1)$. Therefore, $\pi'$ also satisfies obedient. Moreover, the Sender's payoff from $\pi'$ is greater than from $\pi^*$, because $u_S$ is supermodular.

Now we can iterate the process until $d = 0$, and we construct an outcome distribution which is comonotone, obedient, and gives the Sender a higher payoff than $\pi^*$. Since the Sender's payoff from $\pi^*$ is strictly greater than any no-information outcome, the Sender benefits from credible persuasion.

$\square$

## A.5 Proof of Proposition 3

From Theorem 1 of Mensch (2021), if both $u_S$ and $u_R$ are supermodular and $|A| = 2$, there exists a KG optimal outcome distribution that is comonotone. Then by Theorem 1 and Lemma 1, such an outcome distribution is stable. Moreover, if in addition $u_S$ is strictly supermodular, any KG optimal outcome distribution is comonotone. So any KG optimal outcome distribution is stable.

## A.6 Proof of Proposition 4

For each buyers' belief over quality, $\mu \in \Delta(\Theta)$, let $\underline{\theta}_\mu$ denote the smallest $\theta$ in the support of $\mu$. In addition, for each $\mu \in \Delta(\Theta)$, let $\phi_\mu(x) \equiv E_\mu[v(\theta)|\theta \leq x]$ denote the corresponding expected value to buyers when the quality threshold is $\theta \leq x$.[17] Clearly, $\phi_\mu(\cdot)$ is increasing and $\phi_\mu(1) = E_\mu[v(\theta)]$.

**Lemma 4.** *For every $\mu \in \Delta(\Theta)$, there exists a largest fixed point $\theta_\mu^* \in (\underline{\theta}_\mu, 1)$ such that $\phi_\mu(\theta_\mu^*) = \theta_\mu^*$. Moreover, for any $\theta \in (\theta_\mu^*, 1]$, $\phi_\mu(\theta) < \theta$.*

*Proof.* Since $\phi_\mu(\underline{\theta}_\mu) = v(\underline{\theta}_\mu) > \underline{\theta}_\mu$, $\phi_\mu(1) = E_\mu[v(\theta)] < 1$, and $\phi_\mu(\cdot)$ is increasing, from Tarski's fixed point theorem, there exists a largest fixed point $\theta_\mu^* \in (\underline{\theta}_\mu, 1)$ such that $\phi_\mu(\theta_\mu^*) = \theta_\mu^*$. To see the second statement, suppose there exists $\theta \in (\theta_\mu^*, 1)$ such that $\phi_\mu(\theta) \geq \theta$, again from Tarski's fixed point theorem, there exists a fixed point $\theta' \in (\theta_\mu^*, 1)$, which contradicts to $\theta_\mu^*$ being the largest fixed point. $\square$

**Lemma 5.** *Let $\lambda \in \Delta(\Theta \times M)$ be a test, and for every $m \in M$ let $\mu_m \in \Delta(\Theta)$ denote the buyers' posterior belief after observing message $m$. The following strategy profile is a BNE in the game $\langle G, \lambda \rangle$: $\alpha_S(\theta, m) = \theta$, $\beta_1(m) = \beta_2(m) = \theta_{\mu_m}^*$.*

*Proof.* For every message $m$, since $\phi_{\mu_m}(\theta_{\mu_m}^*) = \theta_{\mu_m}^*$, each buyer's expected payoff is 0. Any deviation to a lower bid also gives a payoff of zero. From Lemma 4, for any $\theta \in (\theta_{\mu_m}^*, 1]$, $\phi_{\mu_m}(\theta) < \theta$, so any deviation to a bid higher than $\theta_{\mu_m}^*$ would lead to a negative payoff. Therefore no buyer has incentive to deviate. $\square$

**Lemma 6.** *Let $(\lambda^*, \sigma^*)$ a WD-IC profile. For each message $m$, let $p(m) \equiv \max\{\beta_1^*(m), \beta_2^*(m)\}$ denote the equilibrium market price in the game $\langle G, \lambda^* \rangle$. Then $\phi_{\mu_m}(p(m)) = p(m)$ for each $m \in M$.*

*Proof.* Suppose $\phi_{\mu_m}(p(m)) < p(m)$, then the winning buyer's payoff is negative, and can profitably deviate to bid 0. Now suppose $\phi_{\mu_m}(p(m)) > p(m)$, we show that at least one buyer has an incentive to bid a higher price.

---

[17]For $x$ less than $\underline{\theta}_\mu$ we set $\phi_\mu(x) = v(\underline{\theta}_\mu)$.

If $\beta_1^*(m) \neq \beta_2^*(m)$, then the losing bidder can profitably deviate. Since $\phi_{\mu_m}(\cdot)$ is increasing, there exists small enough $\varepsilon$ such that $\phi_{\mu_m}(p(m) + \varepsilon) > p(m) + \varepsilon$. So the losing bidder can deviate to bidding $p(m) + \varepsilon$ and receives a strictly positive payoff.

If $\beta_1^*(m) = \beta_2^*(m) = b$ for some $b$, we show that both buyers have an incentive to deviate. Let $K \equiv \phi_{\mu_m}(b) - b > 0$. Since ties are broken evenly, each buyer's payoff is $\frac{1}{2}P_{\mu_m}(\theta \leq b)K$. By letting $\varepsilon < \frac{K}{2}$, we have

$$\phi_{\mu_m}(b + \varepsilon) - b - \varepsilon \geq \phi_{\mu_m}(b) - b - \varepsilon = K - \varepsilon > \frac{K}{2}.$$

So if either of the bidders deviates to bidding $b + \varepsilon$, he receives a payoff of $P_{\mu_m}(\theta \leq b + \varepsilon)[\phi_{\mu_m}(b + \varepsilon) - b - \varepsilon] > \frac{1}{2}P_{\mu_m}(\theta \leq b)K$, which is profitable. $\qquad\square$

**Lemma 7.** *If a profile $(\lambda^*, \sigma^*)$ is credible and WD-IC, then there exists a set $E \subset \Theta \times M$ such that $\lambda^*(E) = 1$, and for any $(\theta, m)$, $(\theta', m') \in E$ with $\theta > \theta'$ and $p(m) > p(m')$,*

$$\max\{\theta, p(m)\} + \max\{\theta', p(m')\} \geq \max\{\theta, p(m')\} + \max\{\theta', p(m)\}.$$

*Proof.* Since $(\lambda^*, \sigma^*)$ is WD-IC, trade only happens when the seller's ask price $\alpha^*(\theta, m) = \theta$ is higher than the prevailing market price $p(m) = \max\{\beta_1^*(m), \beta_2^*(m)\}$. The seller's payoff function can therefore be simplified as

$$u_S(\theta, \sigma^*(\theta, m)) = u_S(\theta, \alpha^*(\theta, m), \beta_1^*(m), \beta_2^*(m)) = \max\{\theta, p(m)\}.$$

Recall that credibility requires

$$\lambda \in \underset{\lambda' \in \Lambda(\mu_0, \lambda_M)}{\arg\max} \int u_S(\theta, \sigma^*(\theta, m)) \, d\lambda'(\theta, m).$$

Let $v_S(\theta, m) \equiv u_S(\theta, \sigma^*(\theta, m)) = \max\{\theta, p(m)\}$. From Theorem 1 of Beiglböck, Goldstern, Maresch, and Schachermayer (2009), $\lambda$ is $v_S$-cyclically monotone. That is, there exists a set $E \subset \Theta \times M$ such that $\lambda^*(E) = 1$, and for any sequence $(\theta_k, m_k)_{k=1}^n \in E$,

$$\sum_{k=1}^n v_S(\theta_k, m_k) \geq \sum_{k=1}^n v_S(\theta_k, m_{k+1}).$$

Suppose $(\theta, m), (\theta', m') \in E$, $\theta > \theta'$, and $p(m) > p(m')$. Then $v_S$-cyclical monotonicity implies that

$$v_S(\theta, m) + v_S(\theta', m') \geq v_S(\theta, m') + v_S(\theta', m),$$

which is

$$\max\{\theta, p(m)\} + \max\{\theta', p(m')\} \geq \max\{\theta, p(m')\} + \max\{\theta', p(m)\}.$$

$\square$

In light of Lemma 7, for every credible profile $(\lambda^*, \sigma^*)$ we will focus only on pairs $(\theta, m) \in E$. We will use $\text{proj}_M(E) \equiv \{m \in M : (\theta, m) \in E\}$ to denote the projection of $E$ onto the message space.

Let $\underline{p} = \inf\{p(m)|m \in \text{proj}_M(E)\}$ be the lowest trading price across all messages. For each message $m$, let $\Theta(m) = \{\theta : (\theta, m) \in E\}$ be the set of $\theta$ that is matched with $m$.

**Lemma 8.** Let $(\lambda^*, \sigma^*)$ be a credible and WD-IC profile. For every message $\hat{m} \in \text{proj}_M(E)$ such that $\hat{p} \equiv p(\hat{m}) = \max\{\beta_1^*(\hat{m}), \beta_2^*(\hat{m})\} > \underline{p}$, we have $\Theta(\hat{m}) \cap (\underline{p}, \infty) = \emptyset$.

*Proof.* To prove the lemma, suppose by contradiction that there exists $\hat{\theta} \in \Theta(\hat{m}) \cap (\underline{p}, \infty)$. By the definition of $\underline{p}$, there exists $p'$ with $\underline{p} < p' < \hat{\theta}$ such that $p' = p(m')$ for some $m' \in \text{proj}_M(E)$. Since in equilibrium $p' = E_{\mu_{m'}}[v(\theta)|\theta \in \Theta(m') \cap [0, p']]$, there also exists $\theta' \in \Theta(m')$ such that $\theta' < p'$. Since $(\theta', m'), (\hat{\theta}, \hat{m}) \in E$, by Lemma 7, we have

$$\max\{\theta', \hat{p}\} + \max\{\hat{\theta}, p'\} \leq \max\{\theta', p'\} + \max\{\hat{\theta}, \hat{p}\}$$

Since $\theta' < p'$ by construction, we have

$$\max\{\theta', \hat{p}\} + \max\{\hat{\theta}, p'\} \leq p' + \max\{\hat{\theta}, \hat{p}\} \tag{7}$$

Note also that

$$p' + \max\{\hat{\theta}, \hat{p}\} < \hat{p} + \hat{\theta}. \tag{8}$$

The inequality above follows by considering two possibilities for $\max\{\hat{\theta}, \hat{p}\}$: either $\hat{\theta} \geq \hat{p}$, in which case $p' + \max\{\hat{\theta}, \hat{p}\} = p' + \hat{\theta} < \hat{p} + \hat{\theta}$; or $\hat{\theta} < \hat{p}$, in which case $p' + \max\{\hat{\theta}, \hat{p}\} = p' + \hat{p} < \hat{p} + \hat{\theta}$ as well.

Combining (7) and (8), and noticing $\theta' < p' < \hat{p}$ and $p' < \hat{\theta}$ yield

$$\max\{\theta', \hat{p}\} + \max\{\hat{\theta}, p'\} < \hat{p} + \hat{\theta} = \max\{\theta', \hat{p}\} + \max\{\hat{\theta}, p'\}$$

which is a contradiction.

$\square$

*Proof of the Proposition 4.* To prove our result, we first calculate the seller's profit from an arbitrary credible and WD-IC profile $(\lambda^*, \sigma^*)$. We then show that there exists another credible

and WD-IC profile $(\lambda^0, \sigma^0)$, where $\lambda^0$ is a null information structure, that leads to weakly higher profit for the seller.

Recall that seller's payoff function can be written as $u_S(\theta, \sigma^*(\theta, m)) = \max\{\theta, p(m)\}$, so her ex-ante profit is

$$
\int_{\Theta \times M} \max\{\theta, p(m)\} d\lambda^*(\theta, m) = \int_M \int_0^1 \max\{\theta, p(m)\} \, d\lambda^*(\theta|m) d\lambda_M^*(m)
$$
$$
= \int_M \left[ \int_0^{p(m)} p(m) \, d\lambda^*(\theta|m) + \int_{p(m)}^1 \theta \, d\lambda^*(\theta|m) \right] d\lambda_M^*(m)
$$
$$
= \int_M \left[ p(m) P_{\lambda^*(\theta|m)}(\theta \leq p(m)) + \int_{p(m)}^1 \theta \, d\lambda^*(\theta|m) \right] d\lambda_M^*(m)
$$

By Lemma 6, $p(m) = E_{\lambda^*(\theta|m)}[v(\theta)|\theta \leq p(m)]$, so we can write the integral above as

$$
\int_M \left[ E_{\lambda^*(\theta|m)}[v(\theta)|\theta \leq p(m)] P_{\lambda^*(\theta|m)}(\theta \leq p(m)) + \int_{p(m)}^1 \theta \, d\lambda^*(\theta|m) \right] d\lambda_M^*(m)
$$
$$
= \int_M \left[ \int_0^{p(m)} v(\theta) d\lambda^*(\theta|m) + \int_{p(m)}^1 \theta \, d\lambda^*(\theta|m) \right] d\lambda_M^*(m)
$$

By Lemma 8, for every $m \in \text{proj}_M(E)$, if $p(m) > \underline{p}$ then $\Theta(m) \cap (\underline{p}, \infty) = \emptyset$, so the seller's profit from $(\lambda^*, \sigma^*)$ can be further simplified to

$$
\int_M \left[ \int_0^{\underline{p}} v(\theta) d\lambda^*(\theta|m) + \int_{\underline{p}}^1 \theta \, d\lambda^*(\theta|m) \right] d\lambda_M^*(m)
$$
$$
= \int_0^{\underline{p}} v(\theta) d\mu_0(\theta) + \int_{\underline{p}}^1 \theta \, d\mu_0(\theta). \tag{9}
$$

Having calculated the seller's profit from $(\lambda^*, \sigma^*)$, next we will construct another credible and WD-IC profile $(\lambda^0, \sigma^0)$ with a weakly higher profit, where $\lambda_0$ is the null information structure $\mu_0 \times \delta_{m_0}$.

From Lemma 8, for every $m \in \text{proj}_M(E)$,

$$
\phi_{\mu_m}(p(m)) = E_{\mu_m}[v(\theta)|\theta \leq p(m)] = E_{\mu_m}[v(\theta)|\theta \leq \underline{p}] = \phi_{\mu_m}(\underline{p});
$$

in addition, from Lemma 6, $\phi_{\mu_m}(p(m)) = p(m)$ for every message $m \in \text{proj}_M(E)$. Combining these yields

$$
\phi_{\mu_m}(\underline{p}) = \phi_{\mu_m}(p(m)) = p(m) \geq \underline{p}.
$$

Taking expectation over all messages, we have $\phi_{\mu_0}(\underline{p}) \geq \underline{p}$. By Tarski's fixed point theorem,

there exists a largest $p^0 \in [\underline{p}, 1)$ such that the $\phi_{\mu_0}(p^0) = p^0$.

Using a similar argument as that in Lemma 5, the strategy profile $\sigma^0$ where the seller plays her weakly dominant strategy $\alpha^0(\theta, m_0) = \theta$, and buyers play $\beta_1^0(m_0) = \beta_2^0(m_0) = p^0$ is a BNE in the game $\langle G, \lambda^0 \rangle$.

It remains to show that the seller's profit from $(\lambda^0, \sigma^0)$ is weakly higher than that from $(\lambda^*, \sigma^*)$. Under $(\lambda^0, \sigma^0)$ the seller's profit is

$$
\begin{aligned}
\int_0^1 \max\{\theta, p^0\} d\mu_0(\theta) &= \int_0^{p^0} p^0 d\mu_0(\theta) + \int_{p^0}^1 \theta d\mu_0(\theta) \\
&= p^0 P_{\mu_0}(\theta \le p^0) + \int_{p^0}^1 \theta d\mu_0(\theta) \\
&= E_{\mu_0}[v(\theta)|\theta \le p^0] P_{\mu_0}(\theta \le p^0) + \int_{p^0}^1 \theta d\mu_0(\theta) \\
&= \int_0^{p^0} v(\theta) d\mu_0(\theta) + \int_{p^0}^1 \theta d\mu_0(\theta) \quad (10)
\end{aligned}
$$

Comparing (9) and (10), since $p^0 \ge \underline{p}$ and $v(\theta) > \theta$ for all $\theta$, it follows that

$$
\int_0^{p^0} v(\theta) d\mu_0 + \int_{p^0}^1 \theta d\mu_0 \ge \int_0^{\underline{p}} v(\theta) d\mu_0 + \int_{\underline{p}}^1 \theta d\mu_0.
$$

The seller's profit under $(\lambda^0, \sigma^0)$ is therefore weakly higher than that from $(\lambda^*, \sigma^*)$.

$\square$

## A.7  Proof of Proposition 5

The following lemmas will be useful in our proofs.

**Lemma 9.** *The following correspondence*

$$
B(\mu, \nu) \equiv \arg \max_{\lambda \in \Lambda(\mu,\nu)} \sum_{\theta,m} \lambda(\theta, m) u_S(\theta, \sigma(m))
$$

*is upper hemi-continuous with respect to $(\mu, \nu)$. Thus, the value function*

$$
V(\mu, \nu) \equiv \max_{\lambda \in \Lambda(\mu,\nu)} \sum_{\theta,m} \lambda(\theta, m) u_S(\theta, \sigma(m))
$$

*is continuous.*

*Proof.* The first statement follows directly from Theorem 1.50 of Santambrogio (2015). For any sequence $(\lambda_k, \mu_k, \nu_k) \to (\lambda, \mu, \nu)$ so that $\lambda_k \in B(\mu_k, \nu_k)$ for all $k$, we have $\lambda \in B(\mu, \nu)$. Then

41

$V(\mu,\nu) = \sum_{\theta,m} \lambda(\theta,m)u_S(\theta,\sigma(m)) = \lim_{k\to\infty}\sum_{\theta,m}\lambda_k(\theta,m)u_S(\theta,\sigma(m)) = \lim_{k\to\infty}V(\nu_k,\nu_k),$ which proves the second statement. $\qquad\square$

**Lemma 10.** *Suppose $\mu \in \mathcal{F}_\Theta^N$ and $\nu \in \mathcal{F}_M^N$, then the extreme points of $\Lambda(\mu,\nu)$ is contained in $X^N(\mu,\nu)$.*

*Proof.* Consider the set $Y^N(\mu,\nu) = \{f \in \mathbb{R}_+^{|\Theta|\times|M|} : \sum_\theta f(\theta,\cdot) = N\nu(\cdot), \sum_m f(\cdot,m) = N\mu(\cdot)\}$. From Corollary 8.1.4 of Brualdi (2006), the extreme points of $Y^N(\mu,\nu)$ is contained in $Z^N(\mu,\nu) = \{f \in \mathbb{N}^{|\Theta|\times|M|} : \sum_\theta f(\theta,\cdot) = N\nu(\cdot), \sum_m f(\cdot,m) = N\mu(\cdot)\}$. Since $\Lambda(\mu,\nu) = \{\frac{f}{N} : f \in Y^N(\mu,\nu)\}$ and $X^N(\mu,\nu) = \{\frac{f}{N} : f \in Z^N(\mu,\nu)\}$, the extreme pints of $\Lambda(\mu,\nu)$ is contained in $X^N(\mu,\nu)$. $\qquad\square$

**Lemma 11.** *Let $X,Y$ be metric spaces and $\Gamma : X \rightrightarrows Y$ be a correspondence. If $\Gamma$ is upper hemi-continuous at $x_0 \in X$, and $\Gamma(x_0) = \{y_0\}$ for some $y_0 \in Y$, then $\Gamma$ is continuous at $x_0$.*

*Proof.* For any $\varepsilon > 0$, let $B(y_0,\varepsilon) \subseteq Y$ denote the $\varepsilon$-ball centered at $y_0$. We will show that there exists $\delta > 0$ such that for all $|x - x_0| < \delta$, $\Gamma(x) \cap B(y_0,\varepsilon) \neq \emptyset$, which implies that $\Gamma$ is lhc and therefore continuous.

Now since $\Gamma(x) = \{y_0\} \subseteq B(y_0,\varepsilon)$ and $\Gamma$ is uhc at $x_0$, it follows that there exists $\delta > 0$ such that $\Gamma(x) \subseteq B(y_0,\varepsilon)$ for all $|x - x_0| < \delta$, so $\Gamma(x) \cap B(y_0,\varepsilon) \neq \emptyset$ for all $|x - x_0| < \delta$, which completes the proof. $\qquad\square$

*Proof of Proposition 5 statement 1.* First suppose $(\lambda^*,\sigma^*)$ is not credible. Then there exists $\lambda' \in \Lambda(\mu_0,\lambda_M^*)$ (recall $\mu_0$ is the prior distribution on $\Theta$) and $\varepsilon_0 > 0$ such that

$$\sum_{\theta,m}\lambda^*(\theta,m)u_S(\theta,\sigma^*(m)) < \sum_{\theta,m}\lambda'(\theta,m)u_S(\theta,\sigma^*(m)) - \varepsilon_0$$

By continuity, there exists $\varepsilon_1 > 0$ such that for all $|\lambda - \lambda^*| < \varepsilon_1$ we have

$$\sum_{\theta,m}\lambda(\theta,m)u_S(\theta,\sigma^*(m)) < \sum_{\theta,m}\lambda'(\theta,m)u_S(\theta,\sigma^*(m)) - \varepsilon_0/2 \qquad (11)$$

By Lemma 9, there exists $\varepsilon_2 > 0$ such that for all $|\mu - \mu_0| < \varepsilon_2$ and $|\nu - \lambda_M^*| < \varepsilon_2$, there exists $\lambda \in \Lambda(\mu,\nu)$ with

$$\sum_{\theta,m}\lambda(\theta,m)u_S(\theta,\sigma^*(m)) > \sum_{\theta,m}\lambda'(\theta,m)u_S(\theta,\sigma^*(m)) - \varepsilon_0/2. \qquad (12)$$

Moreover, since the receiver is choosing only pure strategies, there exists $\varepsilon_3$ such that for any $\sigma$ where $|\sigma - \sigma^*| < \varepsilon_3$, $\sigma = \sigma^*$.

Now let $\varepsilon = \min\{\varepsilon_1, \frac{\varepsilon_2}{|\Theta|\times|M|}, \varepsilon_3\}$, By assumption, there exists a finite-sample, credible and R-IC profile $(\nu^N, \phi^N, \sigma^N)$ such that $|\nu^N - \lambda_M^*| < \varepsilon$, $|\sigma^N - \sigma| < \varepsilon$ and $P(|\phi^N(F_\Theta^N) - \lambda| < \varepsilon) > 1 - \varepsilon$.

Under such a finite-sample profile, $\sigma^N = \sigma^*$, and there exists $F_\Theta^N \in \mathcal{F}_\Theta^N$, realized with positive probability, such that $\tilde{\lambda}^* = \phi^N(F_\Theta^N)$ satisfies $|\tilde{\lambda}^* - \lambda^*| < \min\{\varepsilon_1, \frac{\varepsilon_2}{|\Theta|\times|M|}\}$.

Now since $|\tilde{\lambda}^* - \lambda^*| < \varepsilon_1$, by (11) we know that

$$\sum_{\theta,m} \tilde{\lambda}^*(\theta, m)u_S(\theta, \sigma^*(m)) < \sum_{\theta,m} \lambda'(\theta, m)u_S(\theta, \sigma^*(m)) - \varepsilon_0/2 \tag{13}$$

In addition, since $|\tilde{\lambda}^* - \lambda^*| < \frac{\varepsilon_2}{|\Theta|\times|M|}$, we know $F_\Theta^N = \tilde{\lambda}_\Theta^*$ satisfies $|F_\Theta^N - \mu_0| < \varepsilon_2$, and $\nu^N = \tilde{\lambda}_M^*$ satisfies $|\nu^N - \lambda_M^*| < \varepsilon_2$, so by (12) there exists $\tilde{\lambda}' \in \Lambda(F_\Theta^N, \nu^N)$ such that

$$\sum_{\theta,m} \tilde{\lambda}'(\theta, m)u_S(\theta, \sigma^*(m)) > \sum_{\theta,m} \lambda'(\theta, m)u_S(\theta, \sigma^*(m)) - \varepsilon_0/2 \tag{14}$$

Combining (13) and (14), we have

$$\sum_{\theta,m} \tilde{\lambda}'(\theta, m)u_S(\theta, \sigma^*(m)) > \sum_{\theta,m} \tilde{\lambda}^*(\theta, m)u_S(\theta, \sigma^*(m)),$$

Note that $\tilde{\lambda}' \in \Lambda(F_\Theta^N, \nu^N)$, but by Lemma 10, we can replace $\tilde{\lambda}'$ with an extreme point in $X^N(F_\Theta^N, \nu^N)$, and the above inequality still holds. That is, there exists $\hat{\lambda}' \in X^N(F_\Theta^N, \nu^N)$ such that

$$\sum_{\theta,m} \hat{\lambda}'(\theta, m)u_S(\theta, \sigma^*(m)) > \sum_{\theta,m} \tilde{\lambda}^*(\theta, m)u_S(\theta, \sigma^*(m)),$$

Notice that $\tilde{\lambda}^*$ and $\hat{\lambda}'$ are both in $X^N(F_\Theta^N, \nu^N)$, which is a contradiction since by the credibility of $(\nu^N, \phi^N, \sigma^N)$

$$\tilde{\lambda}^* = \phi^N(F_\Theta^N) = \underset{\lambda \in X^N(F_\Theta^N, \nu^N)}{\arg\max} \sum_{\theta,m} \lambda(\theta, m)u_S(\theta, \sigma^*(m)).$$

Second, suppose $(\lambda^*, \sigma^*)$ violates R-IC. Then there exists $\sigma'$ such that

$$\sum_{\theta,m} \lambda^*(\theta, m)u_R(\theta, \sigma'(m)) > \sum_{\theta,m} \lambda^*(\theta, m)u_R(\theta, \sigma^*(m))$$

By continuity, there exist $\eta > 0$ and $\varepsilon_4 > 0$ such that for all $\lambda'$ satisfying $|\lambda^* - \lambda'| < \varepsilon_4$, we have

$$\sum_{\theta,m} \lambda'(\theta, m)u_R(\theta, \sigma'(m)) - \sum_{\theta,m} \lambda'(\theta, m)u_R(\theta, \sigma^*(m)) \geq \eta > 0$$

Let $d \equiv \max_{\theta,a} u_R(\theta, a) - \min_{\theta,a} u_R(\theta, a)$ denote the gap between the Receiver's highest

43

and lowest payoffs. Let $\varepsilon_5 \leq \frac{\eta}{d+\eta}$ and $\varepsilon = \min\{\varepsilon_3, \varepsilon_4, \varepsilon_5\}$. By assumption, there exists a credible and R-IC finite-sample profile $(\nu^N, \phi^N, \sigma^N)$ such that $Pr(|\phi^N(F_\Theta^N) - \lambda^*| \leq \varepsilon) > 1 - \varepsilon$, and $\sigma^N = \sigma^*$. We will show that in the finite sample profile $(\nu^N, \phi^N, \sigma^N)$, the Receiver can profitably deviate from $\sigma^N = \sigma^*$ to $\sigma'$, which contradicts $(\nu^N, \phi^N, \sigma^N)$ being R-IC.

By choosing $\sigma^*$ the Receiver obtains payoff

$$\sum_{F_\Theta^N \in \mathcal{F}_\Theta^N} P^N(F_\Theta^N) \sum_{\theta,m} \phi^N(\theta, m | F_\Theta^N) u_R(\theta, \sigma^*(m))$$

By contrast, the Receiver obtains

$$\sum_{F_\Theta^N \in \mathcal{F}_\Theta^N} P^N(F_\Theta^N) \sum_{\theta,m} \phi^N(\theta, m | F_\Theta^N) u_R(\theta, \sigma'(m))$$

from choosing $\sigma'$. Denote $E^N \equiv \{F_\Theta^N : |\phi^N(F_\Theta^N) - \lambda^*| \leq \delta\}$ so $Pr(E^N) > 1 - \varepsilon$. By switching from $\sigma^*$ to $\sigma'$, the Receiver obtains an extra payoff of

$$\sum_{F_\Theta^N \in \mathcal{F}_\Theta^N} P^N(F_\Theta^N) \left[ \sum_{\theta,m} \phi^N(\theta, m | F_\Theta^N) u_R(\theta, \sigma^*(m)) - \sum_{\theta,m} \phi^N(\theta, m | F_\Theta^N) u_R(\theta, \sigma'(m)) \right]$$

$$= \sum_{F_\Theta^N \in E^N} P^N(F_\Theta^N) \left[ \sum_{\theta,m} \phi^N(\theta, m | F_\Theta^N) u_R(\theta, \sigma^*(m)) - \sum_{\theta,m} \phi^N(\theta, m | F_\Theta^N) u_R(\theta, \sigma'(m)) \right]$$

$$+ \sum_{F_\Theta^N \notin E^N} P^N(F_\Theta^N) \left[ \sum_{\theta,m} \phi^N(\theta, m | F_\Theta^N) u_R(\theta, \sigma^*(m)) - \sum_{\theta,m} \phi^N(\theta, m | F_\Theta^N) u_R(\theta, \sigma'(m)) \right]$$

Note that $\sum_{\theta,m} \phi^N(\theta, m | F_\Theta^N) u_R(\theta, \sigma^*(m)) - \sum_{\theta,m} \phi^N(\theta, m | F_\Theta^N) u_R(\theta, \sigma'(m)) \geq \eta$ for all $F_\Theta^N \in E^N$, while for all $F_\Theta^N \notin E^N$, $\sum_{\theta,m} \phi^N(\theta, m | F_\Theta^N) u_R(\theta, \sigma^*(m)) - \sum_{\theta,m} \phi^N(\theta, m | F_\Theta^N) u_R(\theta, \sigma'(m)) \geq -d$. Together they imply,

$$\sum_{F_\Theta^N \in \mathcal{F}_\Theta^N} P^N(F_\Theta^N) \left[ \sum_{\theta,m} \phi^N(\theta, m | F_\Theta^N) u_R(\theta, \sigma^*(m)) - \sum_{\theta,m} \phi^N(\theta, m | F_\Theta^N) u_R(\theta, \sigma'(m)) \right]$$

$$\geq \eta P^N(E^N) - d(1 - P^N(E^N)).$$

Since $P^N(E^N) > 1 - \varepsilon$, we have

$$\sum_{F_\Theta^N \in \mathcal{F}_\Theta^N} P^N(F_\Theta^N) \left[ \sum_{\theta,m} \phi^N(\theta, m | F_\Theta^N) u_R(\theta, \sigma^*(m)) - \sum_{\theta,m} \phi^N(\theta, m | F_\Theta^N) u_R(\theta, \sigma'(m)) \right]$$

$$> (1 - \varepsilon)\eta - \varepsilon d = \eta - \varepsilon(\eta + d) \geq 0$$

This contradicts the R-IC of $(\nu^N, \phi^N, \sigma^N)$.

$\square$

*Proof of Proposition 5 statement 2.* For each $N \geq 1$, define $\sigma^N = \sigma^*$, $\nu^N \in \arg\min_{\nu \in \mathcal{F}_M^N} |\lambda_M^* - \nu|$, and $\phi^N : \mathcal{F}_\Theta^N \to \cup_{F_\Theta^N \in \mathcal{F}_\Theta^N} X(F_\Theta^N, \nu^N)$ by

$$\phi(F_\Theta^N) \in \arg\max_{\lambda \in X^N(F_\Theta^N, \nu^N)} \sum_{\theta, m} \lambda(\theta, m) u_S(\theta, \sigma^*(m)).$$

By construction, for every $N$, $(\nu^N, \phi^N, \sigma^N)$ is credible and $|\sigma^N - \sigma^*| = 0$. It remains to show that for every $\varepsilon > 0$, there exists large enough $N$, such that

1. $|\nu^N - \lambda_M^*| < \varepsilon$;

2. $P(|\phi^N(F_\Theta^N) - \lambda^*| < \varepsilon) > 1 - \varepsilon$;

3. $(\nu^N, \phi^N, \sigma^N)$ is R-IC.

From the denseness of rational numbers, we know that $\nu^N \to \lambda_M^*$ as $N \to \infty$ so the first statement follows.

To prove the second statement, note that since $(\lambda^*, \sigma^*)$ is strictly credible, $\lambda^*$ is the unique maximizer to

$$\max_{\lambda \in \Lambda(\mu, \nu)} \sum_{\theta, m} \lambda(\theta, m) u_S(\theta, \sigma^*(m)).$$

From Lemma 9, the best response correspondence $B(\mu, \nu) \equiv \arg\max_{\lambda \in \Lambda(\mu, \nu)} \sum_{\theta, m} \lambda(\theta, m) u_S(\theta_i, \sigma^*(m_j))$ is upper hemi-continuous. Since $B(\mu, \nu) = \{\lambda^*\}$ is a singleton, from Lemma 11, $B$ is continuous at $(\mu, \nu)$. Therefore, there exists $\delta > 0$, so that for any $(\mu', \nu')$ such $|\mu - \mu'| < \delta$ and $|\nu - \nu'| < \delta$, we have $|\lambda' - \lambda^*| < \varepsilon$ for every $\lambda' \in B(\mu', \nu')$.

From the Glivenko–Cantelli theorem, for large $N$, $P(|F_\Theta^N - \mu_0| < \delta) > 1 - \varepsilon$. Pick $N$ large enough so that $P(|F_\Theta^N - \mu_0| < \delta) > 1 - \varepsilon$ and $|\nu^N - \mu_M^*| < \delta$. Follows from the definition of $\phi$ and Lemma 10, $\phi(F_\Theta^N) \in \arg\max_{\lambda \in X^N(F_\Theta^N, \nu^N)} \sum_{\theta, m} \lambda(\theta, m) u_S(\theta, \sigma^*(m)) \subset \arg\max_{\lambda \in \Lambda(F_\Theta^N, \nu^N)} \sum_{\theta, m} \lambda(\theta, m) u_S(\theta, \sigma^*(m))$. So with at least $1 - \varepsilon$ probability, $|\phi(F_\Theta^N) - \lambda^*| < \varepsilon$.

Lastly, we show that $(\nu^N, \phi^N, \sigma^N)$ is R-IC for large $N$. Since $(\lambda^*, \sigma^*)$ is strictly R-IC, for any $\sigma \neq \sigma^*$,

$$\sum_{\theta, m} \lambda^*(\theta, m) u_R(\theta, \sigma^*(m)) > \sum_{\theta, m} \lambda^*(\theta, m) u_R(\theta, \sigma(m)).$$

From continuity, there exists $\eta > 0$ such that for any $\lambda$ such that $|\lambda^* - \lambda| < \varepsilon$,

$$\sum_{\theta, m} \lambda(\theta, m) u_R(\theta, \sigma^*(m)) - \sum_{\theta, m} \lambda(\theta, m) u_R(\theta, \sigma(m)) \geq \eta > 0.$$

As we have shown, for any $\varepsilon > 0$, for large enough $N$, $Pr(|\phi(F_\Theta^N) - \lambda^*| \leq \varepsilon) \geq 1 - \varepsilon$. Pick $\varepsilon \leq \frac{\eta}{d+\eta}$, then follow from the same argument above, we have

$$\sum_{F_\Theta^N \in \mathcal{F}_\Theta^N} P^N(F_\Theta^N) \sum_{\theta, m} \phi(\theta, m | F_\Theta^N) u_R(\theta, \sigma^*(m)) > \sum_{F_\Theta^N \in \mathcal{F}_\Theta^N} P^N(F_\Theta^N) \sum_{\theta, m} \phi(\theta, m | F_\Theta^N) u_R(\theta, \sigma(m))$$

for any $\sigma \neq \sigma^*$. So $(\nu^N, \phi^N, \sigma^N)$ is R-IC.

$\square$

## A.8   Proof of Proposition 6

*Proof.* The "only if" direction: Suppose $(\lambda, \sigma)$ is a pure-strategy SPNE outcome of the extensive-form game induced by a pure-strategy SPNE $(\lambda, \rho)$. Then under $(\lambda, \rho)$, the Sender should have no profitable deviation to any other $\lambda' \in \Lambda$ and in particular, any $\lambda'$ that maintains the same marginal distribution on $M$. So for every $\lambda' \in \Lambda$ such that $\lambda'_M = \lambda_M$, we have

$$\sum_\theta \sum_m \lambda(\theta, m) u_S(\theta, \rho(\lambda_M, m)) \geq \sum_\theta \sum_m \lambda'(\theta, m) u_S(\theta, \rho(\lambda'_M, m)).$$

Since $\rho(\lambda_M, m) = \rho(\lambda'_M, m) = \sigma(m)$ for all $m \in M$, the equation above becomes

$$\sum_\theta \sum_m \lambda(\theta, m) u_S(\theta, \sigma(m)) \geq \sum_\theta \sum_m \lambda'(\theta, m) u_S(\theta, \sigma(m))$$

for all $\lambda' \in \Lambda$ such that $\lambda'_M = \lambda_M$. This is the definition of a credible profile in Definition 1.

In addition, the Receiver should have no profitable deviation at every information set $(\lambda_M, m)$ on the equilibrium path. So for every $\rho' \in \Xi$ and every $m \in M$ such that $\lambda_M(m) > 0$, we have

$$\sum_\theta \lambda(\theta, m) u_R(\theta, \rho(\lambda_M, m)) \geq \sum_\theta \lambda(\theta, m) u_R(\theta, \rho'(\lambda_M, m)).$$

Note that $\sigma(m) = \rho(\lambda_M, m)$ for all $m \in M$, so by summing over all $m \in M$, we have

$$\sum_\theta \sum_m \lambda(\theta, m) u_R(\theta, \sigma(m)) \geq \sum_\theta \sum_m \lambda(\theta, m) u_R(\theta, \sigma'(m))$$

for all $\sigma' \in \Sigma$. This is the definition of a R-IC profile in Definition 2. Therefore, the profile $(\lambda, \sigma)$ is both credible and R-IC.

Moreover, notice that after the Sender chooses the uninformative test $\lambda^\circ = \mu_0 \times \delta_{m^\circ}$ for some $m^\circ$, the information set $(\delta_{m^\circ}, m^\circ)$ forms the initial node of a proper subgame. Subgame-perfection at this subgame then requires the Receiver to choose an action $a \in A_0$. Since the Sender always has the option to choose $\lambda^\circ$, her equilibrium payoff cannot be less than

$\min_{a \in A_0} \sum_\theta \mu_0(\theta) u_S(\theta, a)$, so $\sum_\theta \sum_m \lambda(\theta, m) u_S(\theta, \sigma(m)) \geq \min_{a \in A_0} \sum_\theta \mu_0(\theta) u_S(\theta, a)$.

The "if" direction: Suppose that the profile $(\lambda, \sigma)$ is credible and suppose that the Sender's payoff from this profile is greater than the lowest possible no-information payoff. Consider the strategy profile $(\lambda, \rho)$ where $\rho(\lambda_M, m) = \sigma(m)$ for all $m \in M$, but for every $\lambda'_M \neq \lambda_M$, $\rho(\lambda'_M, m) \in \arg\min_{a \in A_0} \sum_\theta \mu_0(\theta) u_S(\theta, a)$ for all $m \in M$. That is, the Receiver chooses the worst (w.r.t. the Sender's payoff) best response to prior belief after observing an off-path marginal distribution on $M$. We claim that $(\lambda, \rho)$ is an SPNE.

We first show that $\rho$ best responds to $\lambda$ in every subgame. We start with the extensive-form game itself. Note that from R-IC, the Receiver best responds to his on-path information sets $(\lambda_M, m)$; that is,

$$\sum_\theta \lambda(\theta, m) u_R(\theta, \rho(\lambda_M, m)) \geq \sum_\theta \lambda(\theta, m) u_R(\theta, \rho'(\lambda_M, m)) \text{ for all } \rho' \in \Xi$$

at every $\lambda_M, m$ such that $\lambda_M(m) > 0$. So $\rho$ best responds to $\lambda$ in the extensive-form game itself.

Next we consider proper subgames of the extensive-form game. Note that among all information sets, the only ones that form the initial node of a proper subgame are those in the form of $(\delta_{m^\circ}, m^\circ)$, which is induced by the Sender choosing an uninformative test $\lambda^\circ = \mu_0 \times \delta_{m^\circ}$ for some $m^\circ \in M$. By our construction, the Receiver chooses the worst (w.r.t. the Sender's payoff) best response to prior belief, so the Receiver's strategy is a best response in these proper subgames. Lastly, for any off-path information set $(\lambda'_M, m)$ that does not define a subgame, SPE has no requirement on the Receiver's strategy. So $\rho$ best responds to $\lambda$ in every subgame.

We now turn to the Sender's strategy $\lambda$ and show that it best responds to $\rho$. From the credibility of $(\lambda, \sigma)$, the Sender has no incentive to deviate to any $\lambda'$ with $\lambda'_M = \lambda_M$. For any other deviation, her payoff is $\min_{a \in A_0} \sum_\theta \mu_0(\theta) u_S(\theta, a_0)$, which is her lowest no-information payoff. Since $\sum_\theta \sum_m \lambda(\theta, m) u_S(\theta, \sigma(m)) \geq \min_{a \in A_0} \sum_\theta \mu_0(\theta) u_S(\theta, a)$, such deviations are not profitable. Therefore $\lambda$ best responds to $\rho$, so $(\lambda, \rho)$ is a pure-strategy SPNE and $(\lambda, \sigma)$ is the corresponding pure-strategy SPNE outcome. $\qquad\square$

# B   Credible Persuasion in Games

In this section we generalize the framework in Section 2.1 to a setting with multiple Receivers, where the Sender can also take actions after information is disclosed. We also allow the state space and action space to be infinite.

Consider an environment with a single Sender (she) and $r$ Receivers (each of whom is a

he). The Sender has action set $A_S$ while each Receiver $i \in \{1, \ldots, r\}$ has action set $A_i$. Let $A = A_S \times A_1 \times \ldots \times A_r$ denote the set of action profiles. Each player has payoff function $u_i : \Theta \times A \to \mathbb{R}$, $i = S, 1, \ldots, r$, respectively. The state space $\Theta$ and action spaces $A_i$ are Polish spaces endowed with their respective Borel sigma-algebras. Players hold full-support common prior $\mu_0 \in \Delta(\Theta)$. We refer to $G = (\Theta, \mu_0, A_S, u_S, \{A_i\}_{i=1}^r, \{u_i\}_{i=1}^r)$ as the base game.

Let $M$ be a Polish space that contains $A$. The Sender chooses a test $\lambda \in \Delta(\Theta \times M)$ where $\lambda_\Theta = \mu_0$: note that this formulation implies that the test generates *public* messages observed by all Receivers. Together the test and the base game constitute a Bayesian game $\mathcal{G} = \langle G, \lambda \rangle$, where:[18]

1. At the beginning of the game a state-message pair $(\theta, m)$ is drawn from the test $\lambda$;

2. The Sender observes $(\theta, m)$ while the Receivers observe only $m$; and

3. All players choose an action simultaneously.

A strategy profile $\sigma : \Theta \times M \to A$ in $\mathcal{G}$ consists of a Sender's strategy $\sigma_S : \Theta \times M \to A_S$ and Receivers' strategies $\sigma_i : M \to A_i$, $i = 1, \ldots, r$. For each profile of Sender's test and players' strategies $(\lambda, \sigma)$, players' expected payoffs are given by

$$U_i(\lambda, \sigma) = \int_{\Theta \times M} u_i(\theta, \sigma(\theta, m)) \, d\lambda(\theta, m) \quad \text{for } i = S, 1, \ldots, r.$$

We now generalize the notion of credibility and incentive compatibility in Section 2 to the current setting. For each $\lambda$, let $D(\lambda) \equiv \{\lambda' \in \Delta(\Theta \times M) : \lambda'_\Theta = \mu_0, \lambda'_M = \lambda_M\}$ denote the set of tests that induce the same distribution of messages as $\lambda$. Definition 4 is analogous to Definition 1, which requires that given the players' strategy profile, no deviation in $D(\lambda)$ can be profitable for the Sender.

**Definition 4.** *A profile $(\lambda, \sigma)$ is **credible** if*

$$\lambda \in \arg\max_{\lambda' \in D(\lambda)} \int u_S(\theta, \sigma(\theta, m)) \, d\lambda'(\theta, m). \tag{15}$$

In addition, Definition 5 generalizes Definition 2, and requires players' strategies to form a Bayesian Nash equilibrium of the game $\langle G, \lambda \rangle$.

**Definition 5.** *A profile $(\lambda, \sigma)$ is **incentive compatible** (IC) if $\sigma$ is a Bayesian Nash equilibrium in $\mathcal{G} = \langle G, \lambda \rangle$. That is,*

$$\sigma_S \in \arg\max_{\sigma'_S : \Theta \times M \to A_S} U_S(\lambda, \sigma'_S, \sigma_{-S}) \quad \text{and} \quad \sigma_i \in \arg\max_{\sigma'_i : M \to A_i} U_i(\lambda, \sigma'_i, \sigma_{-i}) \text{ for } i = 1, \ldots, r. \tag{16}$$

---

[18]The test $\lambda$ can be viewed as "additional information" observed by both the Sender and the Receivers, on top of the base information structure where the Sender observes the state and the Receivers do not observe any signal.

48

Note that in Definition 4, when the Sender deviates to a different test, say $\lambda'$, we use the original strategy profile $\sigma(\theta, m)$ to predict players' actions in the ensuing Bayesian game $\langle G, \lambda' \rangle$. One might worry that the Sender may simultaneously change not only her test but also her strategy $\sigma_S(\theta, m)$ in $\langle G, \lambda' \rangle$. This, however, is unnecessary since the Sender's optimal strategy in $\langle G, \lambda' \rangle$ will remain unchanged: the Sender knows $\theta$ perfectly, her best response in $\langle G, \lambda' \rangle$ depends only on $\theta$ and the Receivers' actions (and not on her own test).