# Credible Persuasion*

Xiao Lin† Ce Liu‡

November 10, 2021

**click here for the latest version**

**Abstract**

We propose a new notion of credibility for information design. 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 characterize when credibility considerations completely shut down informative communication, as well as settings where the sender is guaranteed to benefit from credible persuasion. We apply our results to the market for lemons and bank runs. In the market for lemons, we 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. In the context of bank runs, whether the regulator can credibly perform a stress test to forestall a bank run depends on the welfare cost of a liquidity crisis.

*We are indebted to Nageeb Ali for his continuing guidance and support. We have also benefited from comments and suggestions by Ben Bushong, Carl Davidson, Henrique de Oliveira, Miaomiao Dong, Jon Eguia, Nima Haghpanah, Marc Henry, Tetsuya Hoshino, Yuhta Ishii, Vijay Krishna, Arijit Mukherjee, Harry Pei, Ran Shorrer, Ron Siegel, Jia Xiang, and Hanzhe Zhang.

†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. For example, in the context of schools, if the school 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 gives each student an A without being detected. Analogously, even if a credit rating agency would like to tamper with its rating schemes, any change such tampering induces in the 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.

We ask whether the Sender can persuade the Receiver by using credible disclosure policies. We find that in many settings, credibility can shut down the possibility for persuasion altogether. 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).

We now illustrate our framework with two examples. First, consider a buyer (Receiver) who is choosing 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. To illustrate

| | 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

our results, we assume that all cars are sold at an exogenously fixed price.[1] The payoffs in this example are in Table 1. 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 keeping 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%.

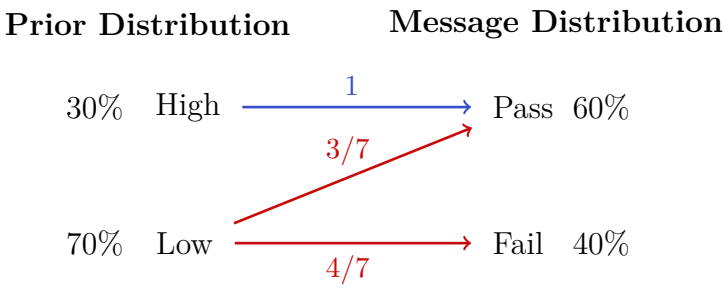



Figure 1: Optimal Commitment Policy

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

[1] In Section 2.5 we study a competitive market for lemons with endogenous prices, and emerge with similar findings.

(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.

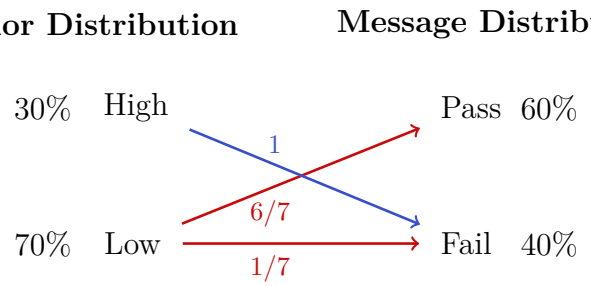



Figure 2: An Undetectable Deviation

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

[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.

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.

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). 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 |
|---|---|---|
| High | 2 | 0 |
| Low | 1 | 0 |

School

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

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 marginal incentive to trade while the Receiver's marginal incentive is stronger; in the school example, by contrast, both the Sender and Receiver have a stronger marginal 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), credibility completely shuts down communication. When players' preferences share the same modularity, the Sender does not always benefit from credible persuasion 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 4.2, we discuss the connection of Theorem 1 to Rochet (1987).

We apply our results to two settings: the market for lemons and bank runs. In the market for lemons, it is well-known that market outcome 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. In another application, we consider a stylized model where a regulator designs a stress test to persuade two large institutional investors to pledge their funds to a bank and forestall a bank run. We show that whether the regulator can credibly disclose information depends on the welfare cost of a liquidity crisis. High welfare costs destroy the credibility of the regulator's stress tests.

An ancillary contribution of this paper is to offer foundations for studying Bayesian persuasion in a large number of settings, which include 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.

The rest of the paper is organized as follows: Section 2 introduces our credibility notion and results first in the setting of a Sender persuading a single Receiver. We then extend our framework to a more general model with multiple Receivers, while also allowing the Sender to take actions. This permits us to apply the model to more applications and in

particular, markets for lemons where prices are determined by actions taken by the seller and multiple buyers. Section 3 discusses two applications: the market for lemons and stress testing for banks. Section 4 provides an extensive-form foundation for our credibility notion, and discusses how the $u_S$-cyclical monotonicity characterization of credibility relates to similar conditions that arise from implementation with transfers. 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.[3] In 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 model 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. A different strand of the literature studies the role of repeated interaction between a patient Sender and a sequence of short-lived Receivers. Best and Quigley (2020) considers how coarse feedback of past realizations of states can substitute for commitment; Mathevet, Pearce, and Stacchetti (2019) allows for the possibility of non-strategic commitment types; Pei (2020) characterizes when Sender's persistent private information about lying cost allows her to achieve her full-commitment payoff.

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. Although in the same spirit, our approach's focus is different in that we study persuasion problems where the Sender's deviations are undetectable if they do not alter the message

[3] Also see Rayo and Segal (2010).

distribution, and characterize information structures where the Sender has no profitable and undetectable deviation. In this way, our credibility notion also connects with the study of quota mechanisms in Jackson and Sonnenschein (2007), Frankel (2014), and Ishii (2016). In both settings, a restriction is placed on the set of feasible deviations. In the context of quota mechanisms, the designer restricts the agent's possible deviations by imposing constraints on the distribution of the agent's reported types (i.e., reporting quotas). In our setting, the Sender's deviations are limited by the distribution of her messages.

Finally, our results offer a plausible foundation for monotone persuasion, which has been the focus of a recent literature. The cyclical monotonicity condition in Theorem 1 reduces to standard monotonicity when the Sender's payoff is supermodular. Monotone information structures have attracted much attention 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.

# 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, $\lambda_\Theta = \mu_0$.[4] The Receiver chooses an action after observing each message according to a pure strategy $\sigma : M \to A$.[5]

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

[4]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.

[5]We focus on pure strategies to abstract from the Receiver using randomization to deter the Sender's deviations, but our analysis can be generalized to allow for mixed strategies. In Section 2.5, we allow the Receiver's action space $A$ to be infinite, which can incorporate mixtures over pure actions.

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*

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

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*

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

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.

Some discussion of our modeling approach is in order. We model the observability of the Sender's message distribution as a partial commitment device. In Section 4.1, 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), the extensive-form analysis rationalizes 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.

**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 $\pi(\Theta, a) > 0$,*

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

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 stability.

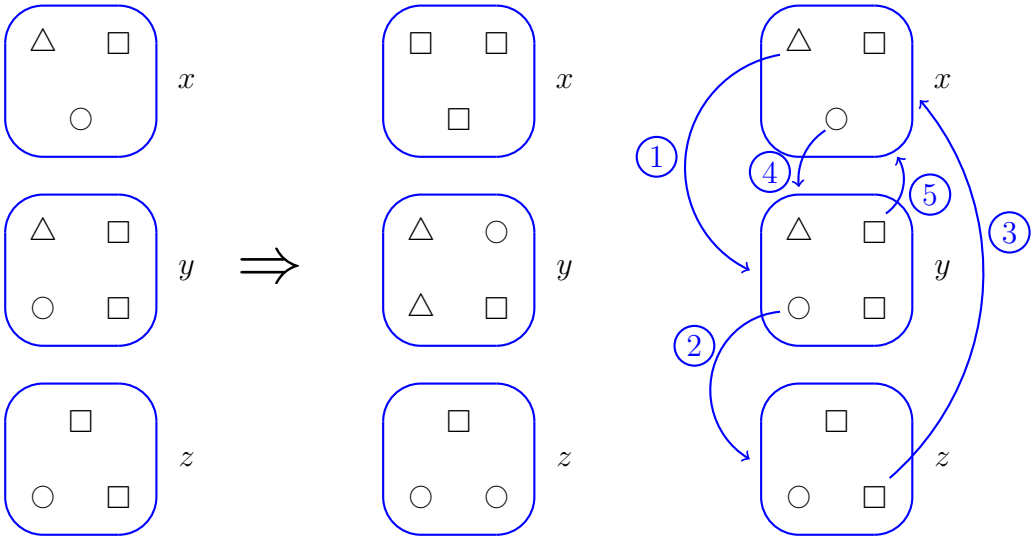
To understand the cyclical monotonicity condition, consider an outcome distribution $\pi$ and a sequence $(\theta_i, a_i)_{i=1}^n$ in the support of $\pi$. A "cyclical" deviation in this case consists

of subtracting $\varepsilon$ mass from $(\theta_i, a_i)$ while adding it to $(\theta_i, a_{i+1})$ for each $n = 1, \ldots, n$, 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.




Outcome distribution $\pi$ A deviation $\pi'$ A cyclical deviation from $\pi$ to $\pi'$

Figure 3: Cyclical Deviation

For sufficiency, the intuition is that any deviation that keeps the marginal distribution on messages unchanged can either be expressed as or approximated by a convex combination of cyclical deviations. To see the idea, let us take a 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 is represented by 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 obvious from the graph that every deviation that involves moving integer numbers of $\square$, $\bigcirc$, or $\triangle$ around can always be written as a cyclical deviation. What is perhaps less obvious are deviations that involve fractions of shapes. We show in the proof that by using the Birkhoff-von Neumann theorem, these deviations can be either expressed as or approximated by cyclical deviations. Therefore, checking cyclical deviations is sufficient for credibility.

## 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). 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

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

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 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 $\theta \geq \theta'$ and $a \geq a'$, 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.

**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 that which we saw in the example: if two distinct messages resulted in different actions from the Receiver, the Sender and Receiver have diametrically opposed preferences as 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.

One might expect that credibility does not limit the Sender's ability to persuade the Receiver when their marginal incentives are aligned. Perhaps surprisingly, this may be false without further assumptions. We illustrate this point using Example 1 in Appendix C. In that example, both the Sender and Receiver have supermodular payoffs. The Sender benefits from

persuasion when she can fully commit to her disclosure policy. However, no stable outcome distribution can give her a higher payoff than the best no-information outcome.

**When the Sender Benefits from Credible Persuasion:** We impose sufficient conditions that guarantee that the Sender benefits from credible persuasion. Let $\overline{a} \equiv \max A$ and $\underline{a} \equiv \min A$ denote the highest and lowest Receiver actions, and let $\overline{\theta} \equiv \max \Theta$ 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 $u_S(\theta, \overline{a}) > u_S(\theta, a)$ for all $\theta$ and $a \neq \overline{a}$, then for generic priors,[6] 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 $u_S(\overline{\theta}, \overline{a}) > u_S(\overline{\theta}, a)$ for all $a \neq \overline{a}$ 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 the patients favor an 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 are quadratic loss preferences as commonly used in models of communication (e.g. Crawford and Sobel, 1982).

The first two parts of Proposition 2 rely 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 $\overline{a}$ is a dominant action for the Sender, and the Sender can benefit from persuasion (under full commitment), then $\overline{a}$ 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 $\delta_{\overline{\theta}}$ and some other posterior $\tilde{\mu}$ that is close to $\mu_0$. At $\delta_{\overline{\theta}}$, the Receiver is induced to choose $\overline{a}$ 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

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

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 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 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.

## 2.5 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 $k$ Receivers (each of whom is a he). The Sender has action set $A_S$ while each Receiver $i \in \{1, \ldots, k\}$ has action set $A_i$. Let $A = A_S \times A_1 \times \ldots \times A_k$ denote the set of action profiles. Each player has payoff function $u_i : \Theta \times A \to \mathbb{R}$, $i = S, 1, \ldots, k$, 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 $G = \left(\Theta, \mu_0, A_S, u_S, \{A_i\}_{i=1}^k, \{u_i\}_{i=1}^k\right)$ 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:[7]

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 $\sigma_i : M \to A_i$, $i = 1, \ldots, k$. For each profile of Sender's test and players' strategies $(\lambda, \sigma)$, players' expected payoffs are given by

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

We now generalize the notion of credibility and incentive compatibility in Section 2 to the current setting. For each $\lambda$, let $D(\lambda) \equiv \left\{\lambda' \in \Delta(\Theta \times M) : \lambda'_\Theta = \mu_0, \lambda'_M = \lambda_M\right\}$ 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*

$$\lambda \in \arg\max_{\lambda' \in D(\lambda)} \int u_S(\theta, \sigma(\theta, m))\, d\lambda'(\theta, m). \tag{3}$$

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 equi-*

[7] 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.

*librium in* $\mathcal{G} = \langle G, \lambda \rangle$. *That is,*

$$\sigma_S \in \underset{\sigma'_S:\Theta\times M\to A_S}{\arg\max}\ U_S(\lambda, \sigma'_S, \sigma_{-S}) \quad \text{and} \quad \sigma_i \in \underset{\sigma'_i:M\to A_i}{\arg\max}\, U_i(\lambda, \sigma'_i, \sigma_{-i}) \text{ for } i = 1, \ldots, k. \tag{4}$$

We say a distribution $\pi \in \Delta(\Theta \times A)$ is an outcome distribution if $\pi_\Theta = \mu_0$. A profile $(\lambda, \sigma)$ induces an outcome distribution $\pi$ if for all measurable subsets $X \subset \Theta$ and $Y \subset A$, we have $\pi(X, Y) = \lambda(\sigma^{-1}(Y) \cap (X \times M))$. We say 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 IC.

In Appendix B.2 we provide an extensive-form game foundation and show that our credibility notion corresponds to the Perfect Bayesian Nash equilibria in that game.

For each state $\theta$ and Receivers' action profile $a_{-S}$, we use $v_S(\theta, a_{-S}) \equiv \max_{a_S \in A_S} u_S\big(\theta, a_S, a_{-S}\big)$ to denote the Sender's indirect utility function. The next result generalizes Theorem 1 to characterize stable outcome distributions in games.

**Theorem 2.** *An outcome distribution* $\pi$ *is stable if and only if:*

1. $\pi$ *is obedient: there exists a set* $E \subset \Theta \times A$, *such that* $\pi$ *concentrates on* $E$, *and for every* $(\hat{\theta}, \hat{a}) \in E$, $\pi$ *satisfies both* $u_S$*-obedience*

$$u_S(\hat{\theta}, \hat{a}) \geq u_S(\hat{\theta}, a'_S, \hat{a}_{-S}) \quad \text{for all } a'_S \in A_S,$$

*and* $u_i$*-obedience for each Receiver* $i = 1, \ldots, k$

$$\int_\Theta u_i(\theta, \hat{a}_{-i}, \hat{a}_i) d\pi(\theta|\hat{a}) \geq \int_\Theta u_i(\theta, \hat{a}_{-i}, a'_i) d\pi(\theta|\hat{a}) \quad \text{for all } a'_i \in A_i.$$

2. $\pi$ *is* $v_S-$*cyclically monotone: there exists a set* $E \subset \Theta \times A$, *such that* $\pi$ *concentrates on* $E$, *and for any* $(\theta^1, a^1), \ldots, (\theta^n, a^n) \in E$ *and* $a^{n+1} \equiv a^1$,

$$\sum_{i=1}^n v_S(\theta^i, a^i_{-S}) \geq \sum_{i=1}^n v_S(\theta^i, a^{i+1}_{-S}).$$

The key difference between Theorem 2 and Theorem 1 is that credibility is now characterized by $v_S$-cyclical monotonicity instead of $u_S$-cyclical monotonicity. This reflects the fact the Sender is privately informed about the state, so when she deviates to a different test, she is able to best respond to Receivers' actions in every state in the new outcome distribution. As a result, when computing the Sender's payoff from cyclical deviations, the indirect utility function $v_S$ is used in place of $u_S$.

# 3 Applications

## 3.1 The Market for Lemons

Having introduced the framework for credible disclosure in games, we now apply the setting to study the market for lemons by adapting the formulation in Mas-Colell, Whinston, and Green (1995). In particular, consider a seller who values an asset she owns (say, a car) at $\theta \in \Theta \subset [0,1]$. There are two buyers (1 and 2) both valuing the car at $v(\theta)$ where $v(\theta)$ is increasing. We assume $v(\theta) > \theta$ for all $\theta \in \Theta$ so there is common knowledge of gain from trade. Buyers share a common prior belief $\mu_0 \in \Delta(\Theta)$.

Without information disclosure, it is well-known that markets may lead to inefficient outcomes because of adverse selection (Akerlof, 1970): despite common knowledge of gain from trade, some cars may not be traded. If the seller can commit to a test to persuade the buyers, she can completely solve the market inefficiency by choosing a test that perfectly reveals $\theta$ to the buyers. Since $v(\theta) > \theta$ for all $\theta \in \Theta$, all cars are traded under full disclosure. It also maximizes the seller's profit when the buyer-side market is competitive, because she captures all the surplus from trade. However, we will show that if the buyers can only observe the message distribution of the seller's test, but not how exactly her test is conducted, then the seller cannot credibly disclose any useful information to the buyers.

Below we first describe the base game without information disclosure. We then augment the base game to allow the seller choosing a test to influence the buyers' trading decisions, and show that despite this, no stable outcome distribution can give the seller a higher profit (or be more efficient) compared to the no-information benchmark.

**The Base Game $G$:** The seller and 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 can choose a test $\lambda$ to disclose information to the buyers. 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 $b_i(m)$, while the seller choose an ask price $a_S(\theta, m)$.

We restrict attention to Bayesian Nash equilibria where the seller plays her *weakly dominant strategy* $\sigma_S(\theta, m) = \theta$. As we show in Lemma 5, such equilibria exist in $\langle G, \lambda \rangle$ for every $\lambda$. In these equilibria, buyers' bids satisfy

$$\sigma_1(m) = \sigma_2(m) = E_{\mu_m}[v(\theta)|\theta \leq \sigma_1(m)]$$

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 such Bayesian Nash equilibria in $\langle G, \lambda_0 \rangle$, so $R_0$ represents the highest equilibrium payoff the seller can achieve when providing no information.

**Proposition 4.** *In every stable outcome distribution, 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.[8] 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.

## 3.2 Bank Runs and Credible Stress Tests

Following the financial crisis of 2008, central banks around the world conduct periodic stress tests for financial institutions to assess their ability to withstand future shocks. Since the results of these tests are disclosed publicly, they are also used as information policies aimed at influencing market beliefs. However, the regulator's interests are often not perfectly aligned with those of the investors, so a key question for investors is whether such disclosure can indeed be credible. In this section, we consider a benevolent regulator whose objective is to maximize the investment returns generated by the banking system while minimizing the risk of a liquidity crisis. We model a stylized bank run game between the regulator and two institutional investors, and evaluate the credibility of the regulator's stress tests.

**The Base Game $G$:** A regulator (Sender) designs a stress test that evaluates the solvency of banks, which are parameterized by an unknown state $\theta \in \Theta \subset [0,1]$ with $|\Theta| < \infty$, and publicly communicates the results of the test to the market. Two large institutional investors $i = 1, 2$ (Receivers) decide whether to pledge their funds to a bank. Each investor has two actions $a_i \in \{0, 1\}$, where $a_i = 1$ represents extending a loan to the bank and $a_i = 0$ represents withdrawal. Investors hold a common prior $\mu_0 \in \Delta(\Theta)$. The bank defaults with probability $\phi(\theta)$ unless both investors choose $a_i = 1$, where $\phi(\theta)$ is a decreasing function in $\theta$.

The regulator's payoff is

$$u_S(\theta, a_1, a_2) = \eta(\theta)(a_1 + a_2) - \phi(\theta)(1 - a_1 a_2)L,$$

where $\eta(\theta)$ is the rate of return the bank can generate with its funds. We assume $\eta(\theta)$ is increasing in $\theta$ since the bank can more effectively pursue profitable ventures when its balance sheet is healthy, as opposed to using the funds to meet only its short-term obligations when $\theta$ is close to 0. $L$ is the welfare cost of a liquidity crisis.

[8] 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 $m'$.

The investors' payoffs are as given in Table 3, where $r(\theta)$ is strictly increasing in $\theta$ and satisfies $r(1) > 0$ and $r(0) < 0$. To understand the investors' payoff structure, note that the bank is guaranteed to remain solvent when both investors pledge their funds, so the investors can secure a normalized payoff of 1 from the interest payment on their funds. When an investor withdraws, he is guaranteed zero return regardless of $\theta$. When only one investor remains pledged to the bank, the interest payment outweighs the prospect of a potential default when $\theta = 1$ (so $r(1) > 0$), but not when $\theta = 0$ (so $r(0) < 0$).

| | $a_2 = 1$ | $a_2 = 0$ |
|---|---|---|
| $a_1 = 1$ | $1, 1$ | $r(\theta), 0$ |
| $a_1 = 0$ | $0, r(\theta)$ | $0, 0$ |

Table 3: Investors' Payoffs

**The Game with Disclosure:** The regulator designs a stress test $\lambda$ and publicly discloses the results to the market. The investors form a posterior belief $\mu_m \equiv \lambda(\cdot|m) \in \Delta(\Theta)$ based on the message $m$ from $\lambda$. Together the test $\lambda$ and the base game $G$ defines a Bayesian game $\langle G, \lambda \rangle$. We focus on the pure-strategy Bayesian Nash equilibria in $\langle G, \lambda \rangle$.

Note that if following a message $m$ the induced posterior belief satisfies $E_{\mu_m}[r(\theta)] > 0$, then the investors have a unique Nash equilibrium $a_1 = a_2 = 1$; if $E_{\mu_m}[r(\theta)] \leq 0$, then there are two pure strategy Nash equilibria, $a_1 = a_2 = 1$ and $a_1 = a_2 = 0$. The regulator does not trust the market to follow her recommendations. Instead, she adopts a robust approach and evaluates a test $\lambda$ under the worst-case scenario. So whenever there are multiple equilibria following a message $m$, the regulator expects the investor to play the worst equilibrium $a_1 = a_2 = 0$. We say a profile $(\lambda, \sigma)$ satisfies *adversarial* receiver incentive compatibility if

- $\sigma_1(m) = \sigma_2(m) = 1$ for every $m \in M$ such that $E_{\mu_m}[r(\theta)] > 0$.
- $\sigma_1(m) = \sigma_2(m) = 0$ for every $m \in M$ such that $E_{\mu_m}[r(\theta)] \leq 0$.

An outcome distribution is *adversarial* if it is induced by a profile that satisfies the adversarial IC.

Suppose the investors observe only the distribution of the regulator's stress-testing evaluations, but not the details of how these stress tests are carried out. The next result characterizes how credibility affects the regulator's ability to disclose information about the banks' solvency to the market.

**Proposition 5.** *There exist $\underline{L}, \overline{L} \in (0, \infty)$ such that when $0 \leq L \leq \underline{L}$, the adversarial optimal full commitment outcome is stable; when $L \geq \overline{L}$, the only adversarially stable outcome distribution is a no-information outcome.*

Proposition 5 says that when the welfare loss from a liquidity crisis $L$ is sufficiently low, credibility does not restrict the regulator's optimal stress testing design. However, as the welfare loss $L$ increases, credibility eventually destroys the regulator's ability to disclose information on banks' solvency to the market.

To see why, note that compared to $(0,0)$, the regulator is always better off when the receivers play $(1,1)$. In fact, the regulator's payoff improves by $2\eta(\theta) + \phi(\theta)L$ when the receivers switch from playing $(0,0)$ to $(1,1)$. When $L$ is small, this difference is increasing in $\theta$ so the regulator has a stronger marginal incentive to induce $(1,1)$ when $\theta$ is high, which makes her marginal incentives aligned with those of the investors. However, if the welfare loss from a liquidity crisis $L$ is sufficiently large, the payoff difference $2\eta(\theta) + \phi(\theta)L$ is decreasing in $\theta$. The regulator now has a stronger marginal incentive to induce $(1,1)$ when the bank is likely to be insolvent, so her marginal incentives are now opposed to those of the investors. The result then follows from arguments similar to those for Proposition 1 and Proposition 3.

# 4 Discussion

## 4.1 An Extensive-Form Foundation

Our solution concept analyzes credible persuasion through the lens of a partial-commitment model. The goal of this section is to provide an extensive-form foundation for this formulation. We propose an extensive-form game between the Sender and the Receiver where we formalize the idea that the Receiver observes the marginal distribution of each test. We show that the set of pure-strategy subgame perfect Nash equilibria of this extensive-form game correspond to the set of profiles $(\lambda, \sigma)$ that are credible, R-IC, and give the Sender higher than her worst no-information payoff. We focus in this section on the single-Receiver environment introduced in Section 2.1, but using the same idea, we also provide an extensive-form foundation for the multiple-Receivers setting of Section 2.5 in Appendix B.2.

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). 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*

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.

## 4.2 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 $y' \in Y$,*[9]

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

Our Theorem 1 and Theorem 2 build 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

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

[9] 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 $\pi^*$.

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.

# 5 Conclusion

This paper offers a new notion of credibility for information disclosure. 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 information disclosure 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 taken by players.

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 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 dealership) 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.

Let us also highlight a number of caveats to our work.

In some settings, the Receiver may observe more than the distribution of messages; for example, she may observes 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 important and interesting to understand how different notions of detectability map into different conditions on the outcome distribution.

In multi-agent settings, we have restricted attention to public messages. There is an additional credibility concern when messages can be sent privately: each Receiver may observe his message but does not observe the messages sent to others. In this setting, the Sender may have a motive to deviate to tests that shift the correlation in messages while keeping the marginal distribution of messages unchanged.

Finally, we have analyzed settings in which the underlying payoff relevant state is exogenous. Many settings feature a moral hazard problem in which the underlying state is an action or effort choice made by the Sender (and hence endogenous). One may envision alleviating the Sender's (effort) incentive constraint by having her first commit to a test that discloses information about her effort choice to the Receiver (e.g., a "monitoring structure"). This would be a case where persuasion is used to mitigate moral hazard. Our results suggest that such a use of persuasion is not credible. If the Receiver responds to messages from the chosen test, the Sender could profitably deviate to a test that induces the same marginal distribution of messages independent of her actions. This would allow the Sender to choose the least-costly action and nevertheless benefit from the responsiveness of the Receiver. As with adverse selection, every credible test would then be uninformative. This suggests that credibility considerations may impede the ability for persuasion to mitigate issues of both moral hazard and adverse selection.

# 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.

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.

Chakraborty, Archishman and Rick Harbaugh. 2010. "Persuasion by Cheap Talk." *American Economic Review* 100 (5):2361–82.

Crawford, Vincent P and Joel Sobel. 1982. "Strategic Information Transmission." *Econometrica: Journal of the Econometric Society* :1431–1451.

Durrett, Rick. 2019. *Probability: Theory and Examples*, vol. 49. Cambridge university press.

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

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

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

Ishii, Yuhta. 2016. "Implementation via Quota Mechanisms." Working Paper.

Ivanov, Maxim. 2020. "Optimal Monotone Signals in Bayesian Persuasion mechanisms." *Economic Theory* :1–46.

Jackson, Matthew O and Hugo F Sonnenschein. 2007. "Overcoming incentive constraints by linking decisions 1." *Econometrica* 75 (1):241–257.

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

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).

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.

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. 2019. "Reputation and Information Design." Working Paper.

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

Min, Daehong. 2021. "Bayesian Persuasion under Partial Commitment." *Economic Theory* :1–22.

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

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.

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

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 is a finite version of Theorem 5.10 of Villani (2008), where the more general version is proved by exploiting duality. In the special case of finite-support probability distributions, we present a direct proof using the Birkhoff-von Neuman theorem, which provides better intuition for why cyclical deviations can be viewed as extreme points of all possible deviations.

**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.* $(1 \Rightarrow 2)$: Suppose $\pi^* \in \arg\max_{\pi \in \Pi(p,q)} \sum_{x,y} \pi(x,y)u(x,y)$, then for any $\pi' \in \Pi(p,q)$, $\sum_{x,y} \pi^*(x,y)u(x,y) \geq \sum_{x,y} \pi'(x,y)u(x,y)$. If $\pi^*$ is not $u$-cyclically monotone, there exists $(x_1, y_1), ..., (x_n, y_n) \in \text{supp}(\pi^*)$, such that $\sum_{i=1}^{n} u(x_i, y_i) < \sum_{i=1}^{n} u(x_i, y_{i+1})$. Without loss of generality, we can assume all pairs $(x_1, y_1), ..., (x_n, y_n)$ are distinct, otherwise the sequence can be broke into two shorter sequences. Let $\varepsilon = \min_{i=1,...,n} \pi(x_i, y_i) > 0$. Construct $\hat{\pi}(x_i, y_i) = \pi^*(x_i, y_i) - \varepsilon$, $\hat{\pi}(x_i, y_{i+1}) = \pi^*(x_i, y_i) + \varepsilon$ for $i = 1, ..., n$. For any other $x, y$, $\pi^*(x,y) = \hat{\pi}(x,y)$. Since $\sum_{x \in X} \hat{\pi}(x,y) = \sum_{x \in X} \pi^*(x,y)$ for all $y$ and $\sum_{y \in Y} \hat{\pi}(x,y) = \sum_{y \in Y} \pi^*(x,y)$ for all $x$, we have $\hat{\pi} \in \Pi(p,q)$. But by construction,

$$\sum_{x,y} \hat{\pi}(x,y)u(x,y) - \sum_{x,y} \pi^*(x,y)u(x,y) = \varepsilon[\sum_{i=1}^{n} u(x_i, y_{i+1}) - \sum_{i=1}^{n} u(x_i, y_i)] > 0,$$

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

$(2 \Rightarrow 1)$: We prove the contraposition. Suppose there exists $\pi'$ such that $\sum_{x,y} \pi'(x,y)u(x,y) > \sum_{x,y} \pi^*(x,y)u(x,y)$. We want to show that there exists $(x_1, y_1), ..., (x_n, y_n) \in supp(\pi^*)$ and a permutation $t : \{1, ...n\} \to \{1, ..., n\}$, such that[10]

$$\sum_{i=1}^{n} u(x_i, y_i) < \sum_{i=1}^{n} u(x_i, y_{t(i)}).$$

First we prove that the statement assuming $\pi^*(x,y)$ and $\pi'(x,y)$ are all rational numbers for all $x, y$. Let $N$ be an integer such that $Np(x)$ and $Nq(y)$ are both integers. Let $\{S_x\}_{x \in X}$ be partition of $\{1, ..., N\}$ indexed by $x$ such that $|S_x| = Np(x)$, and similarly $\{T_y\}_{y \in Y}$ be a partition such that $|T_y| = Nq(y)$. Define the following matrix:

$$M_{ij} = N\frac{\pi^*(x,y)}{|S_x||T_y|} \quad \text{if } i \in S_x, j \in T_y.$$

[10]Proving for permutations is sufficient because one can relabel so that $t(i) = i + 1$.

Notice that $M_{ij}$ is doubly stochastic: for any $x$ and any $i \in S_x$

$$\sum_j M_{ij} = \sum_y \left( N \frac{\pi(x,y)}{|S_x||T_y|} \cdot |T_y| \right) = \frac{Np(x)}{|S_x|} = 1.$$

Similarly, for any $y$ and $j \in T_y$, $\sum_i M_{ij} = 1$.

Let $X(i) = \{x | i \in S_x\}$ and $Y(j) = \{x | j \in T_y\}$. Then

$$\sum_{i,j} M_{ij} u(X(i), Y(j)) = N \sum_{x,y} \pi^*(x,y)\, u(x,y)$$

We can similarly define $M'_{ij}$ so that

$$\sum_{i,j} M'_{ij} u(X(i), Y(j)) = N \sum_{x,y} \pi'(x,y)\, u(x,y)$$

Since

$$\sum_{x,y} \pi'(x,y)\, u(x,y) > \sum_{x,y} \pi^*(x,y)\, u(x,y),$$

we have

$$\sum_{i,j} M'_{ij} u(X(i), Y(j)) > \sum_{i,j} M_{ij} u(X(i), Y(j)).$$

From Birkhoff-von Neumann theorem, there exist permutation matrices $P_{ij}$ and $P'_{ij}$ such that

$$\sum_{i,j} P'_{ij} u(X(i), Y(j)) > \sum_{i,j} P_{ij} u(X(i), Y(j)).$$

Notice that a permutation matrix is equivalent to a mapping $t : \{1, ..., n\} \to \{1, ..., n\}$ such that $P_{ij} = 1$ if and only if $j = t(i)$. So $\sum_{i,j} P_{ij} u(X(i), Y(j)) = \sum_i u(X(i), Y(t(i)))$.

Now we consider a sequence $(x_i, a_i)_{i=1}^n = \big(X(i), Y(t(i))\big)_{i=1}^n$. Every element of the sequence is in the support of $\pi^*$ because $P_{ij} = 1$ only if $M_{ij} > 0$. Similarly, let $t'$ denote the permutation matrix $P'_{ij}$. So $\sum_i u(X(i), t'(Y(i))) > \sum_i u(X(i), Y(t(i)))$. Then there exists a permutation $s : \{1, ..., n\} \to \{1, ..., n\}$ such that $s(t(i)) = t'(i)$. Now

$$\sum_{i=1}^n u(x_i, a_i) < \sum_{i=1}^n u(x_i, a_{s(i)})$$

which contradicts to cyclical monotonicity.

Now we consider the case where $\pi^*(x,y)$ or $\pi'(x,y)$ might be irrational number for some $x, y$. Since rational numbers are dense, for any $\varepsilon$, there exists $\tilde{\pi}^*(x,y)$ and $\tilde{\pi}'(x,y)$ that are all rational numbers, and $|\tilde{\pi}'(x,y) - \pi'(x,y)| < \varepsilon$, $|\tilde{\pi}^*(x,y) - \pi^*(x,y)| < \varepsilon$. Pick $\varepsilon$ small enough

so that $\tilde{\pi}$ and $\pi$ has the same support.

Moreover, since

$$N \sum_{x,y} \pi'(x,y)\, u(x,y) > N \sum_{x,y} \pi^*(x,y)\, u(x,y),$$

for $\varepsilon$ small enough,

$$N \sum_{x,y} \tilde{\pi}'(x,y)\, u(x,y) > N \sum_{x,y} \tilde{\pi}^*(x,y)\, u(x,y),$$

Then the rest of the proof follows from the same argument as the rational numbers case.

□

*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.

Recall that $M \supseteq A$, so we can construct a test $(M, \lambda^*)$ by letting $\lambda^*(\theta, m) = \pi$

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 $(\theta_1, a_1), \ldots, (\theta_n, a_n)$ in $\text{supp}(\pi)$ 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 $(\theta_1, m_1), \ldots, (\theta_n, m_n) \in \text{supp}(\lambda^*)$ 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 $(\theta_1, a_1), \ldots, (\theta_n, a_n) \in \text{supp}(\pi)$ 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 $(\theta_i, m_i) \in \text{supp}(\lambda^*)$, so we have a sequence $(\theta_1, m_1), \ldots, (\theta_n, m_n) \in \text{supp}(\lambda^*)$ that satisfies

$$\sum_{i=1}^{n} u_S(\theta_i, \sigma^*(m_i)) < \sum_{i=1}^{n} u_S(\theta_i, \sigma^*(m_{i+1})), \tag{5}$$

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 $\text{supp}(\lambda^*)$, 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 (5). So $\pi$ must be $u_S$-cyclically monotone.

□

## 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 $t(t(k)) < t(k)$. □

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

$$u_S(\theta_{t(k^*)}, a_{t(k^*)}) + u_S(\theta_{k^*}, a_{t(t(k^*))}) \geq u_S(\theta_k, a_{t(k^*)}) + u_S(\theta_{t(k^*)}, a_{t(t(k^*))}). \tag{6}$$

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 (6), 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

$$\sum_{\theta \in I_1} [u_R(\theta, a_1) - u_R(\theta, a_2)] \frac{\pi(\theta, a_1)}{\pi_a(a_1)} \geq 0 \geq \sum_{\theta' \in I_2} [u_R(\theta', a_1) - u_R(\theta', a_2)] \frac{\pi(\theta', a_2)}{\pi_a(a_2)} \tag{7}$$

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

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

So

$$\begin{aligned}\sum_{\theta\in I_1}[u_R(\theta,a_1)-u_R(\theta,a_2)]\frac{\pi(\theta,a_1)}{\pi_a(a_1)} &\leq \max_{\theta\in I_1}\left\{u_R(\theta,a_1)-u_R(\theta,a_2)\right\}\\ &\leq \min_{\theta'\in I_2}\left\{u_R(\theta',a_1)-u_R(\theta',a_2)\right\} \qquad (8)\\ &\leq \sum_{\theta'\in I_2}[u_R(\theta',a_1)-u_R(\theta',a_2)]\frac{\pi(\theta',a_2)}{\pi_a(a_2)}\end{aligned}$$

Combining (7) and (8), 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 messages 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 $a^\circ \neq \overline{a}$, 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

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

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 $\{\overline{a}, a^\circ\}$. 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,

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

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

For $\overline{a}$, 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 $\overline{\mu} \in \Delta(\Theta)$ such that $\overline{a} \in \arg\max_a \sum_\theta \overline{\mu}(\theta) u_R(\theta, a)$. So for every $a' \neq \overline{a}$,

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

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

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

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

above implies

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

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

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

$$\begin{aligned}
\sum_{\theta,a} \pi^\varepsilon(\theta, a) u_S(\theta, a) &= \sum_{\theta \neq \overline{\theta}} \mu_0(\theta) u_S(\theta, a^\circ) + (\mu_0(\overline{\theta}) - \varepsilon) u_S(\overline{\theta}, a^\circ) + \varepsilon u_S(\overline{\theta}, \overline{a}) \\
&> \sum_{\theta \neq \overline{\theta}} \mu_0(\theta) u_S(\theta, a^\circ) + (\mu_0(\overline{\theta}) - \varepsilon) u_S(\overline{\theta}, a^\circ) + \varepsilon u_S(\overline{\theta}, a^\circ) \\
&= \sum_{\theta} \mu_0(\theta) u_S(\theta, a^\circ) = v_0.
\end{aligned}$$

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 $H_{a,a'} = \{\mu \in \Delta(\Theta) | \sum_\theta \mu(\theta)(u_R(\theta, a) - u_R(\theta, a')) = 0\}$ for any $j \neq k$. Since by assumption, $u_R(\cdot, a) - u_R(\cdot, a') \neq \mathbf{0}$, which implies $H_{a,a'}$ is a hyperplane in the space $\Delta(\Theta)$. By the definition of a hyperplane, $H_{a,a'}$ has measure 0, so $\cup_{a \neq a'} H_{a,a'}$ also has measure 0 in $\Delta(\Theta)$.

For any $\mu \in \Delta(\Theta)/\{\cup_{a \in A} P_a\}$, since the maximizer of $\sum_\theta \mu(\theta) u_R(\theta, a)$, 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 $\Delta(\Theta)$.

□

*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 $\mu^\circ \notin P_{\overline{a}}$. We consider the case $\mu^\circ \notin P_{\overline{a}}$, and the other case can be shown symmetrically. Similar as the previous argument, for $\varepsilon$ sufficiently small, consider the outcome distribution $\pi^\varepsilon \in \Delta(\Theta \times A)$:

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

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.

□

*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.

□

## 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 Theorem 2

We first formalize the setup. There is a probability space $(\Omega, \mathcal{F}, P)$, and complete separable metric spaces $\Theta$, $M$, and $A$.

The sender chooses an information structure, which is an integrable $X : \Omega \to \Theta \times M$ that induces a joint distribution $\lambda \in \Delta(\Theta \times M)$ such that the marginal distribution on $\Theta$ is $\mu_0$. Sender's strategy is a measurable function $\sigma_S : \Theta \times M \to A$, and Receiver $i$'s strategy is a measurable function $\sigma_i : M \to A_i$. A strategy profile is $\sigma = (\sigma_S, \sigma_1, ..., \sigma_r) : \Theta \times M \to A_S \times A_1 \times ... \times A_r$. For each $\sigma$, define $Y^\sigma : \Omega \to \Theta \times A$ such that $Y^\sigma(\omega) = (\theta, a)$ iff $X(\omega) = (\theta, m)$ for some $(\theta, m) \in \sigma^{-1}(a)$. The random variable $Y^\sigma$ induces an outcome distribution $\pi \in \Delta(\Theta \times A)$.

The sender's payoff function is $u_S : \Theta \times A \to \mathbb{R}$ and the receiver $i$'s payoff function is $u_i : \Theta \times A \to \mathbb{R}$. Define $U_S^\sigma = u_S \circ Y : \Omega \to \mathbb{R}$ and $U_i^\sigma = u_i \circ Y : \Omega \to \mathbb{R}$ be random variables that representing the players' payoffs., Then sender and receiver $i$'s expected payoff under strategy profile $\sigma$ are $E[U_S^\sigma]$ and $E[U_i^\sigma]$. Let $\mathcal{F}_M$ be the $\sigma$-algebra induced by $M$, and $\mathcal{F}_A$ be the $\sigma$-algebra induced by $A$. Clearly $\mathcal{F}_M$ is finer than $\mathcal{F}_A$, i.e., $\mathcal{F}_A \subset \mathcal{F}_M$.

A strategy profile $\sigma^*$ is a BNE in $\mathcal{G} = \langle G, \lambda^* \rangle$ if and for any $i = S, 1, ..., r$,

$$E[U_i^{\sigma_i, \sigma_{-i}} | \mathcal{F}_M] \geq E[U_i^{\sigma_i', \sigma_{-i}} | \mathcal{F}_M]$$

for any $\sigma_i'$.

Now we are ready to prove the theorem.

($\Leftarrow$): Suppose an outcome distribution $\pi^* \in \Delta(\Theta \times A)$ with $\pi_\Theta = \mu_0$ satisfies $v_S-$cyclically monotone and obedient. Then clearly a message space $M = A$ with $\lambda^* = \pi^*$ and $\sigma^*$ being the identity mapping induces outcome distribution $\pi$. Moreover, the profile $\lambda, \sigma^*$ is trivially a BNE from the obedient constraint. Now from Theorem 5.10 of Villani (2008), $\pi^*$ satisfies $v_S-$cyclically monotone implies

$$\pi^* \in \arg \max_{\pi \in \Pi(\mu_0, \pi_A^*)} \int v_S(\theta, a) d\pi,$$

where $\Pi(\mu_0, \pi_A^*)$ denotes the set of joint distribution on $(\Theta \times A)$ such that the marginal on $\Theta$ and $A$ are $\mu_0$ and $\pi_A^*$ respectively. Since $\sigma^*$ is an identity mapping,

$$\lambda^* \in \arg \max_{\pi \in \Lambda(\mu_0, \lambda_M^*)} \int v_S(\theta, \sigma(m)) d\lambda.$$

Therefore, we have constructed a credible profile $(\lambda^*, \sigma^*)$ which induces $\pi^*$.

($\Rightarrow$): Suppose a credible profile $(\lambda^*, \sigma^*)$ induces an outcome distribution $\pi^*$. Clearly $\pi_\Theta = \mu_0$. We first show that $\pi$ is obedient. Since $\sigma^*$ is a BNE in $\mathcal{G} = \langle G, \lambda^* \rangle$, for any $i = S, 1, ..., r$,

$$E[U_i^{\sigma_i^*, \sigma_{-i}^*} | \mathcal{F}_M] \geq E[U_i^{\sigma_i', \sigma_{-i}^*} | \mathcal{F}_M]$$

for any $\sigma_i'$. By taking conditional expectation w.r.t. $\mathcal{F}_A$ for both side, we have

$$E[E[U_i^{\sigma_i, \sigma_{-i}} | \mathcal{F}_M] | \mathcal{F}_A] \geq E[E[U_i^{\sigma_i', \sigma_{-i}} | \mathcal{F}_M] | \mathcal{F}_A]$$

for any $\sigma_i'$. Since $\mathcal{F}_A \subset \mathcal{F}_M$, $E[E[U|\mathcal{F}_M]|\mathcal{F}_A] = E[U|\mathcal{F}_A]$ for any random variable $U$. (See, e.g., Theorem 4.1.13 of Durrett (2019)). Therefore, for any $i$,

$$E[U_i^{\sigma_i, \sigma_{-i}} | \mathcal{F}_A] \geq E[U_i^{\sigma_i', \sigma_{-i}} | \mathcal{F}_A]$$

for any $\sigma_i'$, which is the obedience constraint.

Moreover, $(\lambda^*, \sigma^*)$ being credible requires

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

From Theorem 5.10 of Villani (2008), $\lambda^*$ is $\tilde{u}-$cyclically monotone, where $\tilde{u}(\theta, m) \equiv u_S(\theta, \sigma_S^*(\theta, m), \sigma_{-S}^*(m))$. That is, there exists a set $E$ such that $(\Theta \times M) \backslash E$ is negligible, and for every sequence $(\theta_1, m_1), ..., (\theta_n, m_n) \in E$,

$$\sum_{i=1}^n u_S(\theta_i, \sigma_S^*(\theta_i, m_i), \sigma_{-S}^*(m_i)) \geq \sum_{i=1}^n u_S(\theta_i, \sigma_S^*(\theta_i, m_{i+1}), \sigma_{-S}^*(m_{i+1})).$$

Consider a set $D \subset \Theta \times A$ such that $(\theta, a) \in D$ if and only if there exists $(\theta, m) \in \sigma^{*-1}(a)$ such that $(\theta, m) \in E$. Since $E$ is a full measure set, so is $D$. Then for any sequence $(\theta_1, a_1), ..., (\theta_n, a_n) \in D$, there exists sequence $(\theta_1, m_1), ..., (\theta_n, m_n) \in E$ such that $a_{S,i} = \sigma_S^*(\theta_i, m_i)$ and $a_{-S,i} = \sigma_{-S}^*(m_i)$. So

$$\begin{aligned} \sum_{i=1}^n u_S(\theta_i, a_{S,i}, a_{-S,i}) &= \sum_{i=1}^n u_S(\theta_i, \sigma_S^*(\theta_i, m_i), \sigma_{-S}^*(m_i)) \\ &\geq \sum_{i=1}^n u_S(\theta_i, \sigma_S^*(\theta_i, m_{i+1}), \sigma_{-S}^*(m_{i+1})) \end{aligned} \tag{9}$$

Notice that from the $u_S-$obedience, we have that with probability 1, $u_S(\theta, a_S, a_{-S}) = \max_{a_S} u_S(\theta, a_S, a_{-S}) = v_S(\theta, a_{-S})$. Similarly from the requirement of BNE, with probability 1, $u_S(\theta, \sigma_S^*(\theta, m), \sigma_{-S}^*(m)) =$

$\max_{a_S} u_S(\theta, a_S, \sigma^*_{-S}(m))$. So (9) can be rewritten as

$$\begin{aligned}\sum_{i=1}^{n} v_S(\theta_i, a_{-S,i}) &= \sum_{i=1}^{n} u_S(\theta_i, a_{S,i}, a_{-S,i}) \\ &\geq \sum_{i=1}^{n} u_S(\theta_i, \sigma^*_S(\theta_i, m_{i+1}), \sigma^*_{-S}(m_{i+1})) \\ &= \sum_{i=1}^{n} v_S(\theta_i, \sigma^*_{-S}(m_{i+1})) \\ &= \sum_{i=1}^{n} v_S(\theta_i, a_{-S,i+1})\end{aligned}$$

which proves that $\pi^*$ is $v_S-$cyclically monotone.

## A.7 Proof of Proposition 4

To simplify notations, for each $\mu \in \Delta(\Theta)$, let $\phi_\mu(x) = E_\mu[v(\theta)|\theta \leq x]$.[11] That is, $\phi_\mu : [0,1] \to \mathbb{R}$ represents buyer's expected value of those goods whose quality is lower than the bid. Clearly, $\phi_\mu$ is increasing and $\phi_\mu(1) = E_\mu[v(\theta)]$.

**Lemma 4.** *For any $\mu$, there exists a largest $\theta^*_\mu \in (\underline{\theta}_\mu, 1)$ such that*

$$\phi_\mu(\theta^*_\mu) = \theta^*_\mu.$$

*Moreover, for any $\theta \in (\theta^*_\mu, 1)$, $\phi_\mu(\theta) < \theta$.*

*Proof.* Since $\phi_\mu(\underline{\theta}_\mu) = v(\underline{\theta}_\mu) > \underline{\theta}_\mu$ and $\phi_\mu(1) = E_\mu[v(\theta)] < 1$, from Tarski's fixed point theorem, there exists a largest fixed point $\theta^*_\mu \in (\underline{\theta}_\mu, 1)$ such that $\phi_\mu(\theta^*_\mu) = \theta^*_\mu$. Suppose there exists $\theta \in (\theta^*_\mu, 1)$ such that $\phi_\mu(\theta^*_\mu) \geq \theta^*_\mu$, 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. □

**Lemma 5.** *For any $\lambda$, a BNE exists in $\langle G, \lambda \rangle$.*

*Proof.* We show that the strategy profile $a_S(\theta, m) = \theta$, $b_1(m) = b_2(m) = \theta^*_{\mu_m}$ forms an equilibrium. 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. □

[11] For $x$ less than $\underline{\theta}_\mu$, the smallest $\theta$ in the support of $\mu$, we define $\phi_\mu(x) = v(\underline{\theta}_\mu)$.

**Lemma 6.** *In any obedient outcome $\pi$,*

$$\phi_{\pi(\cdot|b_1,b_2)}(\max\{b_1,b_2\}) = \max\{b_1,b_2\}$$

*Proof.* Clearly $\phi_{\pi(\cdot|b_1,b_2)}(\max\{b_1,b_2\}) \geq \max\{b_1,b_2\}$, otherwise the winning bidder can profitably deviate to bid $b_i = 0$, which violates the obedient constraint. Now suppose $\phi_{\pi(\cdot|b_1,b_2)}(\max\{b_1,b_2\}) > \max\{b_1,b_2\}$. We will show that at least one buyer has an incentive to bid a higher price, and this violates the obedient constraint.

First if $b_1 \neq b_2$, then the losing bidder can profitably deviate. Since $\phi_{\pi(\cdot|b_1,b_2)}(\cdot)$ is an increasing function, there exists small enough $\varepsilon$ such that $\phi_{\pi(\cdot|b_1,b_2)}(\max\{b_1,b_2\}+\varepsilon) > \max\{b_1,b_2\}+\varepsilon$. So the losing bidder can increase his bid to $\max\{b_1,b_2\}+\varepsilon$ and receive a strictly positive payoff.

If $b_1 = b_2 = b$, then both bidders have an incentive to deviate. Let $K \equiv \phi_{\pi(\cdot|b_1,b_2)}(b) - b$. Each bidder's expected payoff is $\frac{1}{2}P_{\pi(\cdot|b_1,b_2)}(\theta \leq b)K$. By letting $\varepsilon < \frac{K}{2}$, we have

$$\begin{aligned}\phi_{\pi(\cdot|b_1,b_2)}(b+\varepsilon) - b - \varepsilon &\geq \phi_{\pi(\cdot|b_1,b_2)}(b) - b - \varepsilon \\ &= K - \varepsilon \\ &> \frac{K}{2}.\end{aligned}$$

So if either of the bidder deviates to bid $b+\varepsilon$, he receivers an expected payoff of $P_{\pi(\cdot|b_1,b_2)}(\theta \leq b+\varepsilon)[\phi_{\pi(\cdot|b_1,b_2)}(b+\varepsilon) - b - \varepsilon] > \frac{1}{2}P_{\pi(\cdot|b_1,b_2)}(\theta \leq b)K$, so the obedient constraint is violated. Therefore, $\phi_{\pi(\cdot|b_1,b_2)}(\max\{b_1,b_2\}) = \max\{b_1,b_2\}$ in any obedient outcome. □

**Lemma 7.** *For any $\pi$ that is $v_S-$cyclically monotone, for any $(\theta,b_1,b_2),(\theta',b_1',b_2') \in supp(\pi)$ where $\theta > \theta'$ and $\max\{b_1,b_2\} > \max\{b_1',b_2'\}$,*

$$v_S(\theta,b_1,b_2) + v_S(\theta',b_1',b_2') = v_S(\theta',b_1,b_2) + v_S(\theta,b_1',b_2').$$

*Proof.* First notice that $v_S(\theta,b_1,b_2) = \max\{\theta,b_1,b_2\}$ is continuous. So in statement 2 of Theorem 2, the set $E$ can be replaced with the support of $\pi$. So $v_S$-cyclical monotonicity implies for any $(\theta,b_1,b_2),(\theta',b_1',b_2') \in \text{supp}(\pi)$ where $\theta > \theta'$ and $\max\{b_1,b_2\} > \max\{b_1',b_2'\}$,

$$v_S(\theta,b_1,b_2) + v_S(\theta',b_1',b_2') \geq v_S(\theta',b_1,b_2) + v_S(\theta,b_1',b_2').$$

On the other hand, since $\max\{b_1,b_2\} > \max\{b_1',b_2'\}$, $\max\{x,b_1,b_2\} - \max\{x,b_1',b_2'\}$ is weakly decreasing in $x$, so for $\theta > \theta'$, $\max\{\theta,b_1,b_2\} - \max\{\theta,b_1',b_2'\} \leq \max\{\theta',b_1,b_2\} - \max\{\theta',b_1',b_2'\}$,

which is equivalent to

$$v_S(\theta, b_1, b_2) + v_S(\theta', b_1', b_2') \leq v_S(\theta', b_1, b_2) + v_S(\theta, b_1', b_2').$$

Combining the two inequalities, we conclude that

$$v_S(\theta, b_1, b_2) + v_S(\theta', b_1', b_2') = v_S(\theta', b_1, b_2) + v_S(\theta, b_1', b_2').$$

□

Let $\underline{b} \equiv \min\{\max\{b_1, b_2\} : (\theta, b_1, b_2) \in \text{supp}(\pi)\}$ denote the smallest winning bid in the support of the outcome distribution, and for each $b$ in the supoprt, let $\Theta(b) \equiv \{\theta : b = \max\{b_1, b_2\}, (\theta, b_1, b_2) \in \text{supp}(\pi)\}$ denote the states that are associated with a winning price $b$ under $\pi$.

**Lemma 8.** *If $\pi$ is obedient, then $\underline{b} \geq \underline{\theta} \equiv \min \Theta(\underline{b})$.*

*Proof.* Consider any $(b_1, b_2)$ in the support such that $\max\{b_1, b_2\} = \underline{b}$. Then from Lemma 6 and the definition of $\phi$,

$$\underline{b} = \phi_{\pi(\cdot|b_1,b_2)}(\underline{b}) \geq v(\underline{\theta}) \geq \underline{\theta}.$$

□

**Lemma 9.** *Suppose $\pi$ is obedient and $v_S-$cyclically monotone. For any $b > \underline{b}$, $\Theta(b) \cap (\underline{b}, \infty) = \emptyset$.*

*Proof.* Suppose by contradiction that there exists $b > \underline{b}$, $\theta \in \Theta(b)$ so that $\theta > \underline{b}$. Let $\underline{\theta} = \min \Theta(\underline{b}) \leq \underline{b}$ from Lemma 8. Suppose $(\underline{\theta}, \underline{b}_1, \underline{b}_2) \in \text{supp}(\pi)$ such that $\underline{b} = \max\{\underline{b}_1, \underline{b}_2\}$, and $(\theta, b_1, b_2) \in \text{supp}(\pi)$ such that $b = \max\{b_1, b_2\}$.

Then from

$$\begin{aligned} v_S(\underline{\theta}, \underline{b}_1, \underline{b}_2) + v_S(\theta, b_1, b_2) &= \max\{\underline{\theta}, \underline{b}\} + \max\{\theta, b\} \\ &= \underline{b} + \max\{\theta, b\} \\ &< b + \theta \\ &= \max\{\underline{\theta}, b\} + \max\{\theta, \underline{b}\} \\ &= v_S(\underline{\theta}, b_1, b_2) + v_S(\theta, \underline{b}_1, \underline{b}_2). \end{aligned}$$

where the strict inequality holds because both $\underline{b} + \theta < b + \theta$ and $b + \underline{b} < b + \theta$. However, this leads to a contradiction to Lemma 7. □

**Lemma 10.** *Suppose $\pi$ is obedient and $v_S-$cyclically monotone. Then*

$$\phi_{\mu_0}(\underline{b}) \geq \underline{b}.$$

*Proof.* From Lemma 9, for any $b > \underline{b}$, $\Theta(b) \cap (\underline{b}, \infty) = \emptyset$. So for any $b_1, b_2$ in the support,

$$\phi_{\pi(\cdot|b_1,b_2)}(\max\{b_1, b_2\}) = E_{\pi(\cdot|b_1,b_2)}[v(\theta)|\theta \leq \max\{b_1, b_2\}] = E_{\pi(\cdot|b_1,b_2)}[v(\theta)|\theta \leq \underline{b}]$$

So for any $b_1, b_2$ in the support, $E_{\pi(\cdot|b_1,b_2)}[v(\theta)|\theta \leq \underline{b}] = \max\{b_1, b_2\} \geq \underline{b}$. Take expectation over $b_1, b_2$ yields

$$E_{\mu_0}[v(\theta)|\theta \leq \underline{b}] \geq \underline{b}.$$

□

*Proof of Proposition 4.* From Lemma 9, for any stable outcome $\pi$, the seller's value is

$$\int \max\{\theta, b_1, b_2\} d\pi(\theta, b_1, b_2) = \int_0^1 \max\{\theta, \underline{b}\} d\mu_0(\theta).$$

From Lemma 10, $\phi_{\mu_0}(\underline{b}) \geq \underline{b}$. Since $\phi_{\mu_0}(1) < 1$ and $\phi_{\mu_0}(\cdot)$ is an increasing function, Tarski's fixed point theorem implies there exists a largest fixed point $b^* \in (\underline{b}, 1)$ such that $\phi_{\mu_0}(b^*) = b^*$. From the same argument as in the proof of Lemma 5, under the null information structure $\lambda^0 = \mu_0 \times \delta_{\{m_0\}}$, the strategy profile $a_S(\theta, m_0) = \theta$, $b_1(m_0) = b_2(m_0) = b^*$ is an equilibrium.

In this equilibrium, the seller receives a payoff of $\int_0^1 \max\{\theta, b^*\} d\mu_0(\theta)$ in this equilibrium. Therefore,

$$R_0 \geq \int_0^1 \max\{\theta, = b^*\} d\mu_0(\theta) \geq \int_0^1 \max\{\theta, \underline{b}\} d\mu_0(\theta).$$

□

## A.8 Proof of Proposition 5

Notice that only two strategy profiles, $(a_1, a_2) = (1, 1)$ and $(a_1, a_2) = (0, 0)$, can be induced in any equilibrium. From our equilibrium selection rule, for any belief $\mu$ such that $E_\mu[r(\theta)] > 0$, $(1, 1)$ will be played, and otherwise $(0, 0)$ will be played.

This is equivalent to consider a single representative receiver with two actions $\{I, NI\}$, who has payoff function

$$u_R(\theta, a) = \begin{cases} r(\theta) & \text{if } a = I, \\ 0 & \text{if } a = NI. \end{cases}$$

and the Sender's payoff function is

$$u_S(\theta, a) = \begin{cases} [R(\theta) + L\phi(\theta)] - L\phi(\theta) & \text{if } a = I, \\ -L\phi(\theta) & \text{if } a = NI. \end{cases}$$

We define an order on $\{I, NI\}$, such that $NI \prec I$, then the Receiver's payoff is supermodular in $(\theta, a)$. For large enough $L$, $R(\theta) + L\phi(\theta)$ is strictly decreasing, so the Sender's payoff is strictly submodular. From Proposition 1, the only stable outcome distribution is the no-information outcome. For small enough $L$, $R(\theta) + L\phi(\theta)$ is strictly increasing, so the Sender's payoff is strictly supermodular. From Proposition 3, every optimal full-commitment outcome is stable.

# B Extensive-Form Foundations

## B.1 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. □

## B.2 An extensive-form foundation for Section 2.5

Consider a game of two players, a Sender and $k$ Receivers. State space $\Theta$, message space $M$, and action space $A$ are all finite and exogenously given, where $|M| \geq |A|$. Let $\Lambda = \{\lambda \in \Delta(\Theta \times M) | \lambda_\theta = \mu^\circ\}$ and $\Sigma_S = \{\sigma_S : \Theta \times M \to A_S\}$ and $\Sigma_i = \{\sigma_i : M \to A_i\}$ for $i = 1, ..., k$.

The timeline is as follows:

1. The Sender chooses $\lambda \in \Lambda$;
2. The Sender (observes $\lambda$) chooses $\sigma_S \in \Sigma_S$; The Receivers observe $\lambda_M$ and choose $\sigma_i \in \Sigma_i$;

The terminal nodes of this extensive-form game can be represented by the tuple $(\lambda, \sigma_S, \sigma_1, ...\sigma_k)$. To simplify notations, we let $\sigma_R = (\sigma_1, ...\sigma_k)$. Players' payoffs are:

$$U_S(\lambda, \sigma_S, \sigma_R) = \sum_\theta \sum_m \lambda(\theta, m) u_S(\theta, \sigma_S(\theta, m), \sigma_R(m))$$

$$U_R(\lambda, \sigma_S, \sigma_R) = \sum_\theta \sum_m \lambda(\theta, m) u_R(\theta, \sigma_S(\theta, m), \sigma_R(m))$$

Let $\kappa = \{\rho_S : \Delta(\Theta \times M) \to \Sigma_S\}$. The Sender's strategy space is $\Lambda \times \kappa$ and Receiver $i$'s strategy space is $\Xi_i = \{\rho_i : \Delta(M) \to \Sigma_i\}$. We consider the Perfect Bayesian Equilbria (PBE) of this game.[12]

**Proposition 7.** *A tuple $(\lambda, \sigma_S, \sigma_R)$ is a pure-strategy PBE outcome of the extensive-form game if and only if*

1. *$(\lambda, \sigma_S, \sigma_R)$ is credible and IC; that is, $(\lambda, \sigma_S, \sigma_R)$ satisfies (3) and (4);*
2. *The sender's value under $(\lambda, \sigma_S, \sigma_R)$ is greater than her lowest equilibrium payoff under no information.*

*Proof.* "Only if":

Suppose $(\lambda, \sigma_S, \sigma_R)$ is a PBE outcome induced by PBE $(\lambda \times \rho_S, \rho_R)$.

First, the Sender must maximize her expected payoff after information set $\lambda$, so

$$\rho_S(\lambda) \in \arg \max_{\sigma'_S : \Theta \times M \to A} \sum_\theta \sum_m \lambda(\theta, m) u_S(\theta, \sigma'_S(\theta, m), \rho_1(\lambda_M)(m), ..., \rho_k(\lambda_M)(m))$$

Since $\rho_S(\lambda) = \sigma_S$ and $\rho_i(\lambda_M) = \sigma_i$, we have

$$\sigma_S \in \arg \max_{\sigma'_S} U_S(\lambda, \sigma'_S, \sigma_R)$$

[12] Our definition of PBE is the one used in Mas-Colell, Whinston, and Green (1995). That is, a strategy profile is a PBE if it induces a weak PBE in every subgame.

which is the IC condition for the Sender.

Sequential rationality requires the Receiver to maximize his expected payoff given his belief. Since on-path belief follows Bayes' rule, the Receiver correctly believes that the Sender chooses $\lambda$ and $\sigma_S$, so for any $i = 1, ..., k$,

$$\rho_i(\lambda_M) \in \arg\max_{\sigma_i'} \sum_{\theta} \sum_{m} \lambda(\theta, m) u_S(\theta, \sigma_S(\theta, m), \sigma_i'(m), \sigma_{-i}(m))$$

which is the IC condition for the Receivers.

Moreover, the Sender should have no profitable deviation to $\lambda' \times \rho_S'$ such that $\lambda_M' = \lambda_M$ and $\rho_S'(\lambda') = \sigma_S$ for all $\lambda'$. This requires

$$\sum_{\theta} \sum_{m} \lambda(\theta, m) u_S(\theta, \rho_S(\lambda)(\theta, m), \rho_R(\lambda_M)(m)) \geq \sum_{\theta} \sum_{m} \lambda'(\theta, m) u_S(\theta, \sigma_S(\theta, m), \rho_R(\lambda_M)(m))$$

which is the credibility condition.

Next we show that the Sender's value under $(\lambda, \sigma_S, \sigma_R)$ can be no less than the lowest equilibrium payoff under no information. Notice that any $\hat{\lambda} = \mu_0 \times \delta_{\hat{m}}$ for some $\hat{m}$ defines a proper subgame. PBE implies the players' strategies in this subgame must be an equilibrium. So $(\rho_S(\hat{\lambda}), \rho_1(\hat{\lambda}_M), ..., \rho_k(\hat{\lambda}_M))$ must form an equilibrium under no information. Since the sender can always deviate to $\hat{\lambda}$, her equilibrium payoff must be higher than her lowest equilibrium payoff under no information.

"If": Suppose $(\lambda, \sigma_S, \sigma_R)$ is credible and IC, and the Sender's payoff is greater than her lowest equilibrium payoff under no information. We construct a strategy profile $(\lambda \times \rho_S, \rho_1, ..., \rho_k)$ that is a PBE of the extensive form game. Suppose $(a_S, a_1, ..., a_k)$ is the worst equilibrium for the Sender under no information.

Let $\rho_S(\lambda) = \sigma_S$ and $\rho_i(\lambda_M) = \sigma_i$ for $i = 1, ..., k$. For any $\lambda' \neq \lambda$ such that $\lambda_M' = \lambda_M$, let $\rho_S(\lambda) = \sigma_S$. For any $\lambda'$ such that $\lambda_M' \neq \lambda_M$, let $(\rho_S(\lambda')(\theta, m), \rho_1(\lambda_M')(m), ..., \rho_k(\lambda_M')(m)) = (a_S, a_1, ..., a_k)$ for any $\theta, m$.

Under this strategy profile, from credibility, the Sender has no incentive to deviate to any $\lambda'$ such that $\lambda_M' = \lambda$. She also has no incentive to deviate to $\lambda'$ such that $\lambda_M' \neq \lambda_M$, because it at most gives her the lowest equilibrium payoff under no information. At every information set $\lambda'$ for the sender and $\lambda_M'$ for the receiver, each player's strategy is sequentially rational from the IC condition. Therefore, $(\lambda \times \rho_S, \rho_1, ..., \rho_k)$ constructed above is a PBE of the extensive form game. □

# C Omitted Example

**Example 1.** *Consider* $\Theta = \{0, 1\}$ *with prior* $\mu_0 = P(\theta = 1) = 0.7$ *and* $A = \{a_1, a_2, a_3, a_4\}$. *The Receiver's payoffs are*

| | $a_1$ | $a_2$ | $a_3$ | $a_4$ |
|---|---|---|---|---|
| $\theta = 0$ | 1 | 0.8 | 0.6 | 0 |
| $\theta = 1$ | 0 | 0.6 | 0.8 | 1 |

*The Sender's payoffs are*

| | $a_1$ | $a_2$ | $a_3$ | $a_4$ |
|---|---|---|---|---|
| $\theta = 0$ | 0 | 1 | 0.5 | $-1$ |
| $\theta = 1$ | $-1$ | 1 | 1 | 0 |

*Both players' payoffs are strictly supermodular. The Receiver's best response* $\hat{a}(\mu)$ *is* $a_1, a_2, a_3, a_4$ *respectively when* $\mu \in [0, 0.25]$, $[0.25, 0.5]$, $[0.5, 0.75]$,$[0.75, 1]$.

*Using the concavification approach from Kamenica and Gentzkow (2011), the Sender's indirect utility function* $\hat{v}(\mu)$ *can be visualized by the blue lines in Figure 4. The red line depicts the concave envelope, so at* $\mu_0 = 0.7$, *the Sender strictly benefits from persuasion if she can fully commit.*

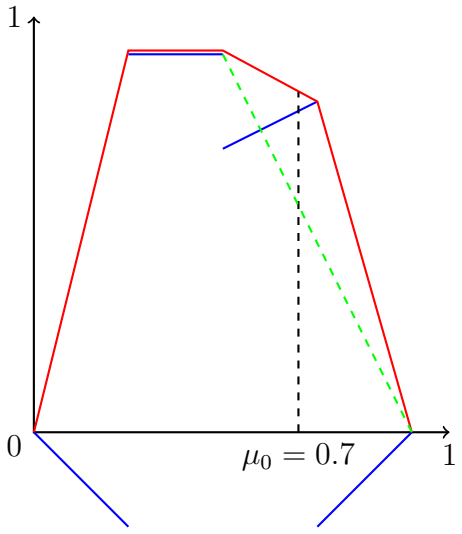



Figure 4: Concavification

*Now consider any stable outcome distribution. Suppose the outcome distribution has at least two actions in the support, then Lemma 1 implies at most one of the actions is matched with more than one state, otherwise comonotonicity is violated. Therefore, in any stable outcome distribution, the induced posterior's support must be included at* $\mu = 0$, $\mu = 1$, *and*

*only one intermediate posterior. For any posterior distribution that induces more than two actions, it can be viewed geometrically that the signal induced by the green dashed line, inducing posteriors* $\mu = 0.5$ *and* $\mu = 1$*, is the optimal one. However, such information structure gives a lower payoff to the sender than the no-information outcome.*