Bayesian Persuasion with Costly Messages[*]

Anh Nguyen[†] Teck Yong Tan[‡]

February 2019

Abstract

We study a model of Bayesian persuasion where the Sender publicly designs a signal structure, *privately* observes the signal realization, and then reports a message to the Receiver at a cost that depends on the signal realization. We provide sufficient conditions for full information revelation by the Sender. These conditions are satisfied under a large class of commonly studied communication games. The persuasion problem then reduces to optimizing over the distribution of the Receiver's posteriors, where each distribution has a cost required to sustain the Sender's credibility in the communication stage. We apply this to study persuasion under partial commitment.

Keywords: Bayesian persuasion, costly messages, partial commitment
JEL Classification: D82, D83, D72, M31, M37

[*] This paper subsumes a previous working paper titled "Bayesian Persuasion under Costly Misrepresentation". We thank Pak Hung Au, Yeon-Koo Che, Navin Kartik, Hongyi Li and Bentley MacLeod for comments and discussions at various stages.

[†] Tepper School of Business, Carnegie Mellon University. Email: anhnguyen@cmu.edu.

[‡] Nanyang Technological University, Singapore. Email: teckyongtan@ntu.edu.sg.

1 Introduction

Many economic situations involve an agent wanting to influence the action of a decision maker. When monetary transfers are not possible, the agent could instead strategically control the decision maker's information to influence her beliefs about the state of the world and thus affect the actions that she takes. Kamenica and Gentzkow (2011) (hereafter, KG) model this as a "Bayesian persuasion" problem whereby the agent (Sender, he) designs an information structure that generates signals about the underlying state to the decision maker (Receiver, she).

In this paper, we study a model of Bayesian persuasion with the innovation that new information is transmitted to the Receiver by the Sender via potentially costly messages. The Sender first publicly commits to a signal structure that generates new information about an unknown state. Upon *privately* observing the signal realization, the Sender sends the Receiver a message at a cost that depends on both the message and signal realization. The Receiver then updates her belief as a Bayesian and takes an action that affects the utility of both players.

Our model adds to the canonical Bayesian persuasion setup in two substantial ways. First, when lying (appropriately defined) is possible but comes at a cost to the Sender, our setup can be interpreted as weakening the Sender's commitment to truthfully revealing all new information to the Receiver, which is a key assumption in the Bayesian persuasion literature. This is of particular relevance to persuasion activities that require expert interpretation or preparation of the new information. For example, drug companies can commit to the scientific research on their new product, but the results require expert interpretation, which is susceptible to misrepresentation. Politicians prepare for re-election by emphasizing areas that they expect to do well in during their term, but the eventual performance statistics can be doctored before they are released from their office.

Second, our model incorporates the Sender's choice of "how" to generate new information under constraints presented at the information transmission stage. One of KG's main insights is that the Bayesian persuasion problem is equivalent to the Sender choosing a distribution of posteriors for the Receiver, subject to the posteriors averaging back to the prior. Each signal from the signal structure is attached to a posterior, and the signals used to encode the respective posteriors are fully interchangeable. Therefore, the possible limitations in the institution for information transmission are neglected. In contrast, in our model, the signal that is used to encode each piece of new information can matter because the Sender's cost to transmit the new information to the Receiver (i.e., message cost) depends on it.

To see why this can matter in practice, consider an example where a drug company is persuading the FDA to approve a new drug, which can be either good or bad. The drug company could design a test that searches for "negative news", such as cases of side effects from the new drug. If the drug is bad, negative news will be found with probability p_B; if the drug is good, negative news is found with only probability $p_G < p_B$. Therefore, when there is news, the FDA's belief about the drug worsens; on the other hand, when there is no news, the FDA's belief improves. Alternatively, the drug company could design a test that searches for "positive news", such as cases of improvement in the health condition after taking the new drug. If the drug is good (respectively, bad), good news will be found with probability q_G (respectively, $q_B < q_G$). Notice that the outcome of the positive-news test has the opposite effect – the FDA's belief about the drug improves (respectively, worsens) when there is news (respectively, no news). However, if the drug company has full flexibility in choosing the probabilities p_G, p_B, q_G and q_B (e.g., through the type of positive or negative news to search for) and is also fully committed to revealing the test outcome, the choice between using a negative-news test or a positive-news test is irrelevant. This is because this is a KG persuasion problem, so it simply reduces to the choice of the distribution of two posteriors.

Suppose now that the FDA has to rely on the drug company to report its test outcome. Moreover, it is infinitely costly for the drug company to fabricate a piece of news, be it positive or negative, but it is always costless to conceal it. In this case, the choice of test becomes important. In a negative-news test, the bad posterior is attached to a signal that is readily manipulated, since the bad news can be costlessly hidden; meanwhile, the good posterior is attached to a signal that the drug company is committed to truthfully report, since the fabrication of news is impossible. Since the drug company always wants to generate the good posterior, its "good message" in the negative-news test is not credible. However, this credibility problem is absent in a positive-news test, since the bad posterior is now attached to a signal that the drug company is committed to truthfully report, while the manipulable signal generates the good posterior. As such, the drug company can still achieve the persuasion outcome of KG despite not having the commitment to truthfully reporting all new information, but it can only be done through a good-news test.

In this paper, we define an *institution* as the triple of the signal space for generating new information, the message space for communicating the new information to the Receiver, and the messaging costs associated with each signal realization. This is distinct from the players' preferences over actions, which has been the focus of the Bayesian persuasion literature. In the example above, the signals available in the institution are verifiable information that cannot be falsified but can be costlessly hidden. Our main model considers more general forms of institutions covering a wide range of communication games that have received great attention in the literature. This allows for flexibility in interpreting what constitutes information misrepresentation by the Sender and its associated costs.

Since new information is privately observed by the Sender and then reported to the Receiver, one immediate question is when there will be full information revelation by the Sender. In equilibrium, the Receiver cannot be systematically "fooled" by the Sender's messages, so the Sender's messaging strategy must satisfy some credibility constraint. As we show in Section 4.1, when messages are costly, the Sender could have an incentive to generate more information than he intends to transmit to the Receiver because this could help him sustain the credibility of a messaging strategy that has a lower ex-ante expected cost. Therefore, we cannot appeal to simple revelation-principle arguments to obtain full information revelation here.

We provide two sufficient conditions whereby if just one of them holds, the Sender-optimal equilibrium is supportable by full information revelation. The first condition is that the Sender has state-independent preferences. The second condition is on the institution, and this condition is satisfied under many common communication protocols, such as communication with verifiable information, costly lying, and cheap talk. Full information revelation in equilibrium implies that if the institution does not permit a piece of new information to be credibly transmitted to the Receiver in the communication stage, then it cannot be generated by the signal structure in the first place.

Armed with this result, we show that the Sender's value from persuasion becomes amenable to the belief-based approach of the Bayesian persuasion literature, whereby the problem reduces to finding the optimal distribution of posteriors held by the Receiver ex-post. The difference is that here, each belief distribution is also attached to a cost that is needed to sustain the credibility of the Sender's messages in the communication game equilibrium. Therefore, if the institution affords the Sender's such credibility costlessly, the Sender can achieve his full commitment persuasion value. Note that this does not always require giving the Sender full commitment to truthfully revealing all information. We provide some natural classes of institutions that provide the required credibility, even though truthful reporting might not be an equilibrium in the associated communication games in isolation.

We then restrict our attention to the institution where the Sender faces a *constant* lying cost $k \geq 0$. This class of institutions allows us to quantify the Sender's commitment level by the value of k. We show that under such institutions, there is no lying in any Sender-optimal equilibrium, but the no-lying constraint restricts the set of feasible signal structures (equivalently, equilibrium distribution of posteriors). As k increases, this set expands in the set inclusion sense, so we can quantify the commitment needed for the Sender to achieve his full commitment persuasion value. Due to the additional messaging costs, the "concavification" method (Aumann and Maschler, 1995; KG) is generally not applicable to our model. However, because the constant lying cost simply puts a constraint on the set of feasible signal structures, we show that this constraint can be represented geometrically when the Sender also has state-independent preferences. Therefore, the optimal distribution of posteriors can still admit a similar geometrical interpretation in this case.

The rest of the paper proceeds as follows. We discuss the related literature in the next section and introduce our model in Section 3. We address issues related to full information revelation in Section 4 and show how the problem reduces to finding the optimal distribution of beliefs with a cost attached to each distribution in Section 5. We focus on the institution with a constant lying cost and study issues related to partial commitment in Section 6. We consider an extension with the Sender also incurring a belief-dependent communication cost in Section 7. Finally, we conclude in Section 8. All omitted proofs are found in Appendix A.

2 Related Literature

Our paper contributes to the literature on information design with a single Sender and Receiver (e.g., KG; Rayo and Segal, 2010).[1] Our main point of departure is to have the Sender privately observe the new information and then send a message to the Receiver. Along this line, Pei (2015) studies a model in which the Sender covertly chooses a signal structure at a cost and then sends a cheap talk message to the Receiver. He shows that the cost causes the Sender to always reveal all information acquired, and the Sender will choose a more informative signal structure than the Receiver would have directly chosen by herself. Argenziano et al. (2016) consider both covert and overt information acquisition within a more restricted set of signal structures, and they make a similar point. In contrast, in our model, the Sender's choice of signal structure is always overt, there is no exogenous cost for the signal structure, and the Sender's message is potentially costly. Therefore, the driving force behind information revelation in our model is very different, and the cost of our signal

[1] See Kamenica (Forthcoming) for a recent survey on this and other areas of information design.

structure is endogenously derived from the communication game.

Gentzkow and Kamenica (2017) study a similar setup as ours; however, they restrict their attention to verifiable disclosure games (*à la* Grossman, 1981; Milgrom, 1981) in the communication stage,[2] and they show that full information revelation is always supportable in the Sender-optimal equilibrium. By contrast, we allow for more general forms of communication games through costly messages,[3] many of which, unlike verifiable disclosure games, do not admit a truthful reporting equilibrium in isolation. The aspect of costly messages also relates our paper to the literature on strategic communication with lying costs (Kartik et al., 2007; Kartik, 2009); the difference is that the Sender's information in this literature is exogenously given.

Our paper is also closely related to some recent papers that weaken the Sender's commitment in Bayesian persuasion problems. Lipnowski and Ravid (2017) (hereafter, LR) study the set of Sender-optimal equilibrium outcomes in cheap talk games when the Sender has state-independent preferences, and they show that this question can be addressed by an information design approach subject to the Sender's incentive constraint to truthfully reveal his information. When framed in this way, this becomes a Bayesian persuasion problem where the Sender has zero commitment to truthfully revealing information. As in LR, the credibility of the Sender's messages is an important issue in our paper; by contrast, full information revelation is not immediate in our model with costly messages and possibly state-dependent Sender preferences (see Section 4.1). Our model admits the setup of LR in the institution with a constant lying cost with $k = 0$. When $k > 0$, the Sender's value from persuasion has a geometrical interpretation that is similar to LR's characterization using quasiconcave envelopes of the Sender's value function.

Relative to LR, Min (2017), Fréchette et al. (2018) and Lipnowski et al. (2018) consider "intermediate" models, where, with some exogenously given probability, the Sender can costlessly lie about the signal realization, and with the complement probability (which measures the Sender's commitment level), the Sender must report truthfully. One implication of their models is that unless the Sender can already achieve his full commitment payoff under zero

[2] However, Gentzkow and Kamenica (2017) also allow for multiple senders and costly signal structures.

[3] KG also considered an extension of their model with messaging costs in their online appendix. However, a "persuasion mechanism" there is the Sender's choice of both signal structure and the messaging cost function. Therefore, having messaging costs that penalize lying by an infinite cost – which KG termed an "honest mechanism" – is always optimal because that essentially gives the Sender full commitment power. Our setup is quite different from that because our messaging costs are exogenously given, so whether a mechanism is honest or not is no longer the choice of the Sender.

commitment,[4] his equilibrium payoff is capped strictly below his full commitment payoff under any partial commitment level. In contrast, in our model, for a large class of institutions, there is a "sufficient" level of commitment, which is still below full commitment, that allows the Sender to achieve his full commitment payoff.

Another paper with the same theme is Guo and Shmaya (2018). To frame it more closely to our context, their model can be interpreted as one where the Sender *privately* chooses a signal structure, privately observes the signal realization and reports the resulting belief over the states at a cost of lying (i.e., how miscalibrated the reported belief is). They show that the Sender is generally better off as the cost intensity increases, and the Sender can achieve his full commitment payoff under a sufficiently high but finite cost intensity. Our model is quite different since our signal structure is publicly known and we focus on costs on messages instead; however, we also consider lying costs with respect to the true beliefs in Section 7. Finally, Best and Quigley (2017) endogenize the Sender's commitment by his reputational concerns; by contrast, the Sender's commitment in our model comes from the institution for information generation and transmission. More broadly, we view our paper as complementary to these works in studying information design under partial commitment. In this respect, our paper is also related to works on mechanism design with limited commitment, e.g., Bester and Strausz (2001); Skreta (2006, 2015); Deb and Said (2015); Doval and Skreta (2018).[5]

Since the new information for persuasion in our model is "signaled" (rather than directly transmitted) to the Receiver, our paper also bears some relation to the voluminous literature on signaling. In contrast to classic signaling games (Spence, 1973), the Sender's private information is endogenously determined here. In and Wright (2017) study a class of endogenous signaling games whereby (at least part of) the Sender's type is privately and deterministically chosen by the Sender. By contrast, our Sender's type is stochastically determined from a publicly chosen distribution. With more direct relations to communication, Austen-Smith and Banks (2000), Kartik (2007) and Karamychev and Visser (2017) allow the Sender to "burn money" to signal information; Fuchs (2015) and Kolotilin and Li (2018) allow the Sender to make voluntary transfers to the Receiver for the same purpose. The difference is that the cost of money burning or voluntary transfers is not directly determined by the Sender's information, whereas the messaging cost is a direct function of the new information in our paper.

[4] Lemma 2 of Best and Quigley (2017) shows that this generically will not happen.

[5] Except for Bester and Strausz (2001), these papers study dynamic mechanism design problems.

3 A Model of Persuasion with Costly Messages

3.1 Setup

There are two players – a Sender (he) and a Receiver (she) – and an unknown state $\omega \in \Omega$. The state space Ω is finite with at least two elements, and the common prior is $\beta^o \in int(\Delta\Omega)$.[6] The game proceeds in two stages. In stage 1, the Sender publicly chooses the signal structure. The set of signal realizations (or simply signals) is S, which satisfies Assumption 1 below, and a signal structure is a measurable map $\pi : \Omega \to \Delta S$, where $\pi(\cdot|\omega) \in \Delta S$ is the probability measure over S in state ω. In stage 2, which is the communication stage, the Sender *privately* observes the signal realization $s \in S$ and then reports a message $m \in M$ to the Receiver, where M is a complete and separable metric space. The Receiver then takes an action $a \in A$, which affects the payoffs of both players.

If the Receiver takes action a in state ω, she obtains utility $u(a, \omega)$, while the utility of the Sender is $v(a, \omega) - c(m|s)$. v is the Sender's *payoff*, which is assumed to be bounded above by some $\bar{v} < \infty$, and $c(m|s) \geq 0$ is the cost of reporting m after observing s.

Assumption 1. *$S = \mathcal{S}^B \times \mathbb{R}$, where $\mathcal{S}^B$ is a complete and separable metric space. For any two signals $s = \left(s^B, x\right)$ and $s' = \left(s^{B'}, x'\right)$, $c(m|s) = c(m|s')\ \forall m \in M$ if $s^B = s^{B'}$.*

We call the set $\mathcal{S}^B$ the *base signal* set, and the base dimension of a signal completely determines its messaging costs. Therefore, we will often abuse notation for the messaging cost function c by also letting it be a mapping from $M \times \mathcal{S}^B$ to $\mathbb{R}^+$ (instead of the original definition of a mapping from $M \times S$ to $\mathbb{R}^+$). The existence of the second dimension is to allow each available base signal, which has a set of associated messaging costs, to be replicated multiple times to generate different information for the Sender if needed.[7] To ignore irrelevant signals, we assume that for every $s \in \mathcal{S}^B$, there exists some $m \in M$ such that $c(m|s) < \infty$ (otherwise, the Sender will never use that signal to generate information).

[6] Throughout, ΔX denotes the set of Borel probability measures of set X, and "*int*" refers to the interior of the set. Therefore, $int(\Delta X)$ refers to the set of Borel probability measures that have full support.

[7] This can be interpreted intuitively as allowing the Sender to have more expertise in learning information than the "establishment" whose perception of the Sender's messages determines his messaging costs. For example, a drug company could test for side effects from its new drugs, and these side effects could have varying degrees of seriousness. However, judging its seriousness requires extra expertise that is not possessed by the public. Therefore, whether the drug company has misrepresented its test result (which determines its messaging cost) is evaluated based on only its report about the presence or absence of a side effect but not on the seriousness of it. In the model, the availability of a side effect is determined by the realization of the base signal dimension, while the degree of seriousness is the second dimension, which does not affect the messaging cost.

We call the triple $\left\{\mathcal{S}^B, M, c\right\}$ the *institution* for information generation and transmission, which is distinct from the players' preferences over actions represented by u and v. A natural class of institutions is one where truth-telling is costless to the Sender – i.e., for every signal $s \in \mathcal{S}^B$, there exists a different message $m \in M$, which is interpreted as the "truth" for s, such that $c\left(m|s\right) = 0$. As noted in the introduction, such a setup can then be viewed as weakening the Sender's commitment in the Bayesian persuasion problem. In particular, the Sender has *full commitment* when his messaging cost is infinite for sending any message but the truth (i.e., the KG setup), and he has *partial commitment* when the messaging cost is finite for some non-truthful messages. As a benchmark, we will use the term "*the Sender's' full commitment payoff*" to denote his value from persuasion under the KG setup given the players' preferences.

3.2 Strategies and equilibrium

The Sender's strategy is a choice of signal structure π and a messaging rule represented by a measurable map $\mu : S \to \Delta M$, where $\mu\left(\cdot|s\right)$ is the probability measure over M after the Sender observes signal s. Since the signal realization is privately observed by the Sender, the Sender and the Receiver can potentially hold different beliefs about the state in equilibrium. To differentiate the players' beliefs, we will use "σ" to denote the Sender's belief and "ρ" to denote the Receiver's belief.[8]

Specifically, under a signal structure π, $\sigma_\pi\left(\cdot|s\right) \in \Delta\Omega$ denotes the Sender's posterior upon observing signal s at the end of stage 1 – i.e., for any $\omega \in \Omega$ and Borel set $\hat{S} \subseteq S$ in the support of π,

$$\pi\left(\hat{S}|\omega\right)\beta^o\left(\omega\right) = \sum_{\omega' \in \Omega} \beta^o\left(\omega'\right) \int_{s \in \hat{S}} \sigma_\pi\left(\omega|s\right) d\pi\left(s|\omega'\right). \tag{1}$$

As for the Receiver, under a signal structure π and a Sender's messaging strategy μ, we denote $\rho_{\pi,\mu}\left(\cdot|m\right) \in \Delta\Omega$ as her posterior upon receiving message m – i.e., for any $\omega \in \Omega$ and Borel set $\hat{M} \subseteq M$ in the support of the joint measure of π and μ,

$$\int_{s \in S} \mu\left(\hat{M}|s\right) d\pi\left(s|\omega\right)\beta^o\left(\omega\right) = \sum_{\omega' \in \Omega} \beta^o\left(\omega'\right) \int_{s \in S}\int_{m \in \hat{M}} \rho_{\pi,\mu}\left(\omega|m\right) d\mu\left(m|s\right) d\pi\left(s|\omega'\right). \tag{2}$$

For convenience, we will define the Receiver's action strategy α as a measurable map from

[8] Since S and M are complete and separable, the regular conditional probabilities (i.e., posterior beliefs) when conditioned on s and m, respectively, exist – see Shiryaev (1996), chapter 7. By the Radon-Nikodym theorem, σ_π and $\rho_{\pi,\mu}$ defined in equations (1) and (2), respectively, define the players' respective posteriors almost everywhere.

her belief (instead of the message sent by the Sender) to actions – i.e., $\alpha : \Delta\Omega \to \Delta A$, where $\alpha(\cdot|\rho) \in \Delta A$ is the probability measure over action set A when the Receiver holds belief ρ about the state. We let $\bar{A}(\rho) := \arg\max_{a \in A} \sum_{\omega\in\Omega} \rho(\omega) u(a,\omega)$ denote the set of Receiver-optimal actions when the Receiver holds belief ρ, and we assume that $\bar{A}(\rho)$ is non-empty for all $\rho \in \Delta\Omega$.

Both players are expected utility maximizers, so we sometimes abuse notation for the players' payoff functions to let them denote expected payoffs as well. This is done by allowing the second arguments of u and v to also be the respective players' beliefs (i.e., let $v(a,\sigma) = \sum_\omega \sigma(\omega) v(a,\omega)$ and $u(a,\rho) = \sum_\omega \rho(\omega) u(a,\omega)$), and the first argument of v to also be a distribution of actions (i.e., let $v(\alpha(\cdot|\rho),\omega) = \int v(a,\omega)\, d\alpha(a|\rho)$). Jointly, this implies that $v(\alpha(\cdot|\rho),\sigma) = \sum_\omega \sigma(\omega) \int v(a,\omega)\, d\alpha(a|\rho)$.

Our equilibrium notion is the perfect Bayesian equilibrium, with a focus on Sender-optimal equilibria.

Definition 1. $(\mu;\alpha)$ is a perfect Bayesian equilibrium (PBE) of the stage-2 communication game under π if

1. Upon receiving message $m \in M$, the Receiver forms posterior $\rho_{\pi,\mu}(\cdot|m) \in \Delta\Omega$ using Bayes' rule according to (2) whenever possible.[9]

2. (Receiver's best response) $\alpha \in \mathcal{A}^*$, where

$$\mathcal{A}^* := \Big\{\ \hat{\alpha} \mid \forall \rho \in \Delta\Omega,\ a \in A \text{ is in the support of } \hat{\alpha}(\cdot|\rho) \implies a \in \bar{A}(\rho) \Big\}. \tag{R-IC}$$

3. (Sender's best response) For any $s \in S$ and $m \in M$ in the support of $\mu(\cdot|s)$,

$$v\Big(\alpha(\cdot|\rho_{\pi,\mu}(\cdot|m)), \sigma_\pi(\cdot|s)\Big) - c(m|s) \geq v\Big(\alpha(\cdot|\rho_{\pi,\mu}(\cdot|m')), \sigma_\pi(\cdot|s)\Big) - c(m'|s) \ \ \forall m' \in M. \tag{S-IC}$$

[9] As usual, the Sender's best response puts restrictions on the Receiver's belief when she receives off-equilibrium messages. To avoid distraction by this issue (which adds no substantial insight here), we assume that there exists a^o in the set of the Receiver's optimal actions at the prior (i.e., $a^o \in \bar{A}(\beta^o)$), with $v(a^o,\sigma) = -\infty\ \forall \sigma \in \Delta\Omega$. Action a^o is interpreted as the Receiver choosing to break away from the relationship with the Sender, which is very costly for the Sender; this is equivalent to assuming that the Sender has a very low outside option which is also state-independent. We then let the Receiver hold the prior belief β^o after every off-equilibrium message and play a^o. In all our subsequent examples, we will not explicitly specify action a^o, with the understanding that such a "leave-the-relationship-action" is always available to the Receiver.

Let

$$V\left(\pi,\mu;\alpha\right):=\sum_{\omega\in\Omega}\beta^{o}\left(\omega\right)\int_{s\in S}\int_{m\in M}v\left(\alpha\left(\cdot|\rho_{\pi,\mu}\left(\cdot|m\right)\right),\sigma_{\pi}\left(\cdot|s\right)\right)d\mu\left(m|s\right)d\pi\left(s|\omega\right) \tag{3}$$

$$C\left(\pi,\mu\right):=\sum_{\omega\in\Omega}\beta^{o}\left(\omega\right)\int_{s\in S}\int_{m\in M}c\left(m|s\right)d\mu\left(m|s\right)d\pi\left(s|\omega\right) \tag{4}$$

V and C are, respectively, the Sender's expected payoff and messaging cost in equilibrium under a strategy profile $(\pi,\mu;\alpha)$.

Definition 2. A *Sender-optimal equilibrium* is a strategy profile $(\pi,\mu;\alpha)$ that maximizes the Sender's ex-ante expected utility

$$W\left(\pi,\mu;\alpha\right):=V\left(\pi,\mu;\alpha\right)-C\left(\pi,\mu\right), \tag{5}$$

subject to $(\mu;\alpha)$ being a PBE of the stage-2 communication game under π, as defined in Definition 1. The Sender's *value from persuasion* is his expected utility in a Sender-optimal equilibrium.

Remark 1. In contrast to most papers in the information design literature, we do not restrict the Receiver to take the Sender-optimal action at each belief (see (R-IC)). This is because such a restriction is not always consistent with a Sender-optimal equilibrium here. This is due to the requirement for the Sender's best response in (S-IC) in equilibrium – it is possible that by not always breaking indifference in favor of the Sender ex-post, the Receiver's action helps to sustain (i.e., satisfies (S-IC)) a more favorable ex-ante distribution of posteriors for the Sender. Therefore, this also implies that we do not have a general "revelation principle" in which the Sender makes an incentive-compatible action recommendation to the Receiver.[10] We illustrate this with an example in Section 6.3.

4 Revelation in the Communication Game

This section studies when the Sender will reveal all his information to the Receiver. At first thought, one might think that full information revelation would follow from a "standard"

[10] Lipnowski and Ravid (2017) also make this point, which also arises in their problem, and relate it to the literature on mechanism design under limited commitment (see cited papers in Section 2). As they very neatly sum it up, in that literature, the principal's information has to be limited due to her limited commitment, so agents cannot always truthfully report their types; here, it is the Sender's influence over the Receiver's actions that requires limitation.

revelation principle argument: if the Sender were to garble information at the communication stage, we can always replicate the equilibrium with a different signal structure that does the associated information garbling on behalf of the Sender. To illustrate that this argument does not hold here when messages are costly, we begin with an example in Section 4.1 where full information revelation does not arise in any Sender-optimal equilibrium. We then provide two general conditions in Section 4.2 for full information revelation to be supportable in a Sender-optimal equilibrium.

4.1 Example without full information revelation

Let $\Omega = \{0, 1\}$ with a uniform prior. The set of base signals is $\mathcal{S}^B = \{x, y, z\}$, and the message set is $M = \{l, h\}$. To simplify notation, we denote the belief as the probability of $\omega = 1$. The Receiver's utility is such that it results in a set of Receiver's optimal actions summarized by

$$\bar{A}(\rho) = \begin{cases} \{1\} & \text{if } \rho = 0.2 \\ \{3\} & \text{if } \rho = 0.8 \\ \{0\} & \text{if } \rho \neq 0.2, 0.8 \end{cases}$$

The Sender's payoff is $v(a, \omega) = a\omega$, and his messaging cost c is as follows:

$$\begin{aligned} c(l|x) &= 0.1 \quad ; \quad c(h|x) = 0 \\ c(l|y) &= 0 \quad ; \quad c(h|y) = 0.1 \\ c(l|z) &= 0.1 \quad ; \quad c(h|z) = \infty \end{aligned}$$

Given $\bar{A}(\rho)$, it is immediate that in any Sender-optimal equilibrium, one message must give the Receiver a belief of 0.2, and the other message must give her a belief of 0.8. We first consider constructing an equilibrium with full information revelation. This implies that after every observed signal, the Sender holds the same belief as the belief attached to the message that he sends to the Receiver under his equilibrium messaging strategy.

Ideally, the Sender would use signals with base signals x and y to generate information and then report h after x and l after y to minimize messaging costs. Consider first attaching belief 0.8 to a signal with base x and belief 0.2 to a signal with base y. If the Receiver expects the Sender to follow the messaging strategy above, her belief is 0.8 after $m = h$, and it is 0.2 after $m = l$. Since belief 0.8 is the Sender's preferred belief for the Receiver, the Sender's credibility to convey this belief is undoubted. However, upon observing the signal with base y, the Sender's belief is 0.2, so his expected utility from sending message l is 0.2, while his expected utility from sending h is $(0.2 \times 3) - 0.1 = 0.5$. Therefore, the Sender's equilibrium constraint (S-IC) is violated at the signal with base y, so this construction is not an equilibrium. By a symmetric argument, attaching belief 0.8 to a signal with base y and belief 0.2 to a signal with base x cannot be sustained as an equilibrium either.

The only way the Sender's equilibrium constraint can be satisfied at belief 0.2 is if this belief is generated by a signal with base z – i.e., let a signal with base x generate a Sender's belief of 0.8 and a signal with base z generate a Sender's belief of 0.2, and the Sender reports h after x and l after z. Abusing notation by denoting a signal by just its base dimension, this set of beliefs is achieved by signal structure π, where

$$\begin{aligned} \pi\left(x|\omega=1\right) &= 0.8 \;\; ; \;\; \pi\left(z|\omega=1\right) = 0.2 \\ \pi\left(x|\omega=0\right) &= 0.2 \;\; ; \;\; \pi\left(z|\omega=0\right) = 0.8 \end{aligned}$$

Since x and z are realized with equal probability, the Sender's expected payoff is $\frac{1}{2}\left(0.8 \times 3\right) + \frac{1}{2}\left(0.2 \times 1\right) = 1.3$, and his expected messaging cost is $\frac{1}{2} \times 0.1 = 0.05$. This gives the Sender his highest expected utility if we restrict the equilibrium to have full information revelation.

Now, consider a different structure $\hat{\pi}$, where (with the same abuse of notation for signals):

$$\begin{array}{lll} \hat{\pi}\left(x|\omega=1\right) = 0.8 \;\; ; & \hat{\pi}\left(y|\omega=1\right) \frac{4}{45} \;\; ; & \hat{\pi}\left(z|\omega=1\right) = \dfrac{1}{9} \\ \hat{\pi}\left(x|\omega=0\right) = 0.2 \;\; ; & \hat{\pi}\left(y|\omega=0\right) = 0.8 \;\; ; & \hat{\pi}\left(z|\omega=0\right) = 0 \end{array}$$

It is readily verified that the Sender's belief upon observing x, y and z are, respectively, 0.8, 0.1 and 1. Let the Sender report $m = h$ after x and report $m = l$ after both y and z (i.e., we do not have full information revelation now). The Receiver's belief upon observing messages h and l will then be, respectively, 0.8 and 0.2, as before. Notice that the Sender's equilibrium constraints (S-IC) are satisfied here. This is immediate for x and z. As for y, the Sender's expected utility from sending $m = h$ is $(0.1 \times 3) - 0.1 = 0.2$, which is the same as his expected utility from sending $m = l$. By noting that the ex-ante probabilities of x, y and z being observed are, respectively, 0.5, $\frac{4}{9}$ and $\frac{1}{18}$, the Sender's expected payoff is $\frac{1}{2}\left(0.8 \times 3\right) + \frac{4}{9}\left(0.1 \times 1\right) + \frac{1}{18}\left(1 \times 1\right) = 1.3$ (as in under π above), while his expected message cost is just $\frac{1}{18} \times 0.1 < 0.05$. Therefore, the Sender's expected utility is higher than under any equilibrium with full information revelation.

To explain the intuition behind this, we first note that the distributions of the Receiver's belief are the same under the two signal structures π and $\hat{\pi}$ (i.e., 0.2 and 0.8 with equal probability). The difference between $\hat{\pi}$ and π is that $\hat{\pi}$ further splits the belief 0.2 in a mean-preserving way to two beliefs for the Sender – namely, beliefs 0.1 and 1. To see why this is helpful, recall that when the Sender holds beliefs 0.2 (which happens only under π), it is not incentive compatible for him to report a message that induces a Receiver's belief of 0.2 if this is at a signal with base y; therefore, a signal with base z, which has a higher messaging cost, must be used. However, incentive compatibility is satisfied if the Sender instead holds a lower belief of 0.1. Therefore, under $\hat{\pi}$, information that will be used to induce a Receiver's belief of 0.2 is partially generated by a signal with base y, which incurs zero messaging cost; on the other hand, under π, it is always generated by a signal with base z, which incurs a messaging cost of 0.1. We can verify that $\hat{\pi}$ is, indeed, the optimal signal structure because it minimizes the use of signals with base z to generate information in general.

4.2 Conditions for full information revelation

In this subsection, we provide two separate conditions on the primitives and show (in Proposition 1 below) that when either one of them holds, the Sender-optimal equilibrium is always supportable by full information revelation. The first condition is on the Sender's preferences, while the second condition is on the institution.

Condition 1. The Sender's preference is state-independent: $v(a,\omega) = v(a,\omega')$ $\forall a \in A$, $\omega, \omega' \in \Omega$.

While this is a substantive restriction, state-independent preferences are satisfied in many situations, which include the two examples in the introduction – the drug company cares only about getting the FDA's approval, and the politician cares only about getting votes.

The second condition is about the institution. Let $\mathcal{M} : \Delta S \rightrightarrows M$, where $\mathcal{M}(\Lambda)$ is the set of messages m' such that $c(m'|\cdot)$ is integrable with respect to the measure $\Lambda \in \Delta S$.[11]

Condition 2. For any $m \in M$ and any measure $\Lambda \in \Delta S$ such that $m \in \mathcal{M}(\Lambda)$, there exists $\bar{s} \in S$ such that

1. $c(m|\bar{s}) \leq \int_{s\in S} c(m|s)\, d\Lambda(s)$.

2. $\forall m' \in \mathcal{M}(\Lambda)\,,\;\; c(m|\bar{s}) - \int_{s\in S} c(m|s)\, d\Lambda(s) \leq c(m'|\bar{s}) - \int_{s\in S} c(m'|s)\, d\Lambda(s)\,.$

3. $\forall m' \in M \backslash \mathcal{M}(\Lambda)$, $\;c(m'|\bar{s}) = \infty$.

[11] i.e., $\mathcal{M}(\Lambda) := \left\{m' \in M \middle| \int_{s\in S} |c(m'|s)|\, d\Lambda(s) < \infty\right\}$.

Condition 2 can be interpreted as every message having a most "efficient signal" to report it. In particular, fixing a message m, we compare the cost of reporting m after its most efficient signal $\bar{s}$ with the expected cost of reporting m after any distribution of signals. Condition 2.1 states that the former must be lower. Condition 2.2 and Condition 2.3 imply that this cost saving from reporting m after its most efficient signal $\bar{s}$ is also greater than using $\bar{s}$ for reporting other message m'. More importantly, the following lemma shows that this condition holds for a large class of institutions that penalize lying.

Lemma 1. *The following institutions satisfy Condition 2:*

- *(Constant lying cost.)* $M = \mathcal{S}^B$*. For any* $m \in M$ *and* $s \in \mathcal{S}^B$, $c\left(m|s\right) = 0$ *if* $m = s$*, and* $c\left(m|s\right) = k \geq 0$ *if* $m \neq s$*.*

- *("Distance" lying cost.)* $M = \mathcal{S}^B$*. For any* $m \in M$ *and* $s \in \mathcal{S}^B$, $c\left(m|s\right) = d\left(m, s\right)$*, where* d *is any metric on* $\mathcal{S}^B$ *(equivalently, on* M*).*[12]

- *(Quadratic lying cost.)* $\mathcal{S}^B$ *is any convex subset of* $\mathbb{R}$ *and* $M = \mathcal{S}^B$*. For any* $m \in M$ *and* $s \in \mathcal{S}^B$, $c\left(m|s\right) = \left(m - s\right)^2$*.*

- *(Partial verifiability.)* $\mathcal{S}^B = \mathcal{V} \cup \mathcal{N}$*, where* $\mathcal{V}$ *and* $\mathcal{N}$ *are disjointed and non-empty sets, and* $M = \mathcal{S}^B \cup \{\phi\}$*. For any* $m \in M$ *and* $s \in \mathcal{S}^B$, $c\left(m|s\right) = 0$ *if* $m = s$, $m = \phi$ *or* $m \in \mathcal{N}$*, and* $c\left(m|s\right) = \infty$ *otherwise.*

- *(Costly disclosure.)* $\mathcal{S}^B = \mathcal{V} \cup \mathcal{N}$*, where* $\mathcal{V}$ *and* $\mathcal{N}$ *are disjointed and non-empty sets, and* $M = \mathcal{S}^B \cup \{\phi\}$*. For any* $m \in M$ *and* $s \in \mathcal{S}^B$, $c\left(m|s\right) = k > 0$ *if* $m = s \in \mathcal{V}$, $c\left(m|s\right) = 0$ *if* $m = \phi$ *or* $m \in \mathcal{N}$*, and* $c\left(m|s\right) = \infty$ *otherwise.*

The institution with a constant lying cost treats all forms of lying as equal, and it also encompasses cheap talk communication (Crawford and Sobel, 1982) when $k = 0$. The institutions with a "distance" lying cost and a quadratic lying cost give intuitive measures of the degree of lying, and the latter has been used in the literature on strategic communication with lying costs (Kartik et al., 2007; Kartik, 2009).

The institutions with partial verifiability and costly disclosure consider communication protocols whereby some signals are hard evidence. $\mathcal{V}$ is the set of verifiable signals (i.e., hard evidence) which are impossible to falsify, $\mathcal{N}$ is a set of non-verifiable signals which the Sender can costlessly claim that he received a signal in this set (e.g., the signals in $\mathcal{N}$ are easily

[12] A metric on a set X is a function $d : X \times X \to [0, \infty)$ such that for any $x, y, z \in X$, (i) $d\left(x, y\right) \geq 0$, (ii) $d\left(x, y\right) = 0$ if and only if $x = y$, (iii) $d\left(x, y\right) = d\left(y, x\right)$, and (iv) $d\left(x, z\right) \leq d\left(x, y\right) + d\left(y, z\right)$.

forged), and ϕ is a null message which is analogous to the Sender not reporting anything. Communication games with verifiable information were first studied by Grossman (1981) and Milgrom (1981) who consider situations where the Sender can be vague about his private information but he cannot tell an outright lie. They show that the Receiver's skepticism can cause information to "unravel", and the Sender fully reveals his private information in equilibrium. In the accounting literature, Dye (1985) and Jung and Kwon (1988) show that the unraveling result breaks down when the Sender can sometimes receive no information, which then allows him to costlessly feign ignorance whenever he wants. Verrecchia (1983) also shows that the Grossman-Milgrom unraveling result can break down when the Sender has to also pay a cost to reveal information, which is analogous to the institution with costly disclosure.

Proposition 1. *Suppose that $(\mu;\alpha)$ is a PBE of the stage-2 communication game under π. If the Sender's preference is state-independent (i.e., Condition 1 holds) $\underline{\text{or}}$ the institution satisfies Condition 2, there exist $\bar{\pi}$ and $\bar{\mu}$ such that*

1. *$\bar{\mu}$ is a pure and fully separating strategy,*[13]
2. *$(\bar{\mu};\alpha)$ is a PBE of the stage-2 communication game under $\bar{\pi}$, and*
3. *$W(\bar{\pi},\bar{\mu};\alpha) \geq W(\pi,\mu;\alpha)$.*

Proposition 1 provides conditions that allow us to restrict our attention to the Sender playing only pure and separating messaging strategies. When this happens, the Receiver continues to perfectly learn all new information generated by the signal structure in equilibrium (i.e., full information revelation), despite the Sender's ability to misrepresent it. We explain the idea behind the proof of Proposition 1 next, and we illustrate how this result helps us solve for the equilibrium in Section 5.

To see why pooling messages always results in a weakly suboptimal equilibrium for the Sender when either of Conditions 1 or 2 holds, let us first suppose that the Sender is already playing a pure messaging strategy, but the Sender pools signals s and s' to message m. Let σ and σ' be, respectively, the Sender's beliefs upon observing s and s', and let the Receiver's belief after receiving m be ρ, which determines the distribution of actions that the Sender faces from reporting m. Since the Receiver correctly conjectures the Sender's messaging strategy in equilibrium, σ and σ' must average out to ρ. Suppose that $c(m|s) < c(m|s')$.

[13] $\bar{\mu}$ is a pure strategy if for every $s \in S$, $\bar{\mu}(\cdot|s)$ is a Dirac measure on some $m \in M$; $\bar{\mu}$ is a fully separating strategy if for any $m \in M$, if m is in the support of both $\mu(\cdot|s)$ and $\mu(\cdot|s')$, then $s = s'$.

Consider another signal structure that shifts the probability of s' (assuming the presence of a mass for simplicity) to s and have the Sender continue to report m after s. The Sender's belief after s is now ρ, the Receiver's belief (and thus actions) after m is unchanged at ρ in equilibrium, and the Sender's expected messaging cost is now lower. When the Sender's preference is state-independent (i.e., Condition 1 holds), his own belief does not affect his payoff. Therefore, his equilibrium constraint (i.e., (S-IC)) at signal s is unchanged, which means that this modified signal structure supports an equilibrium that gives the Sender a higher expected utility.

When the Sender's preference is state-dependent, part of the argument above remains valid. In particular, because σ and σ' will average out to ρ and the expected payoffs are linear in the beliefs, the Sender's expected payoff with a distribution of beliefs over σ and σ' (i.e., in the original signal structure) will be the same as his payoff when he holds belief ρ (i.e., after the modification). However, the equilibrium constraint (S-IC) at signal s, which was satisfied when the Sender had belief σ, may no longer be satisfied when he holds belief ρ. In this case, if Condition 2 holds, we can instead shift the probabilities of both s and s' to a signal $\bar{s}$ pinned down by Condition 2 and have the Sender report m after $\bar{s}$, which still results in a Receiver's belief of ρ after receiving m. The second part of Condition 2, together with the linearity property of payoffs in beliefs, allows the Sender's equilibrium constraint (S-IC) at this new signal $\bar{s}$ to be written as a convex combination of the previous equilibrium constraints of s and s', and it is thus satisfied. The first part of Condition 2 then implies that the Sender's equilibrium expected messaging cost is lower. Jointly, this implies that this modification also supports an equilibrium that gives the Sender a higher expected utility.

Proposition 1 is then obtained by combining the arguments in the preceding two paragraphs with the observation that any randomization over messages by the Sender after any signal realization can be replaced by a signal structure that does the randomization "on behalf" of the Sender while maintaining his equilibrium constraint (S-IC) and the Receiver's posterior at every message. Intuitively, this is because by Assumption 1, for every signal s, there exists another signal s' that has the same base dimension as s and thus has the same messaging costs as s. Therefore, if the Sender's messaging strategy randomizes over messages m and m' after signal s, it is possible to construct another signal structure where some of the conditional probabilities of s are moved to s' according to the Sender's messaging randomization, thus replacing the randomization effect.

Proposition 2. *Suppose that Condition 1 or Condition 2 holds, and a Sender-optimal equilibrium exists. Then there exists a Sender-optimal equilibrium where the numbers of signals* *and messages used in equilibrium are, respectively, less than or equal to* $|\Omega|$.

Proposition 2 is related to the result that only $|\Omega|$ posteriors are needed in the canonical Bayesian persuasion model (see Proposition 4 in the online appendix of KG). While the proof here also uses results in convex analysis,[14] it requires additional care for two reasons. The standard argument for this result in the literature is to consider a set of $|\Omega|$-dimensional elements, each representing a belief (which is of dimension $|\Omega| - 1$) and the Sender's utility under that belief, and then show that the element consisting of the prior and the Sender's ex-ante expected utility is in the convex hull of that set; thus, it can be written as a convex combination of $|\Omega|$ or less elements from that set. The first complication in our model arises from quantifying the Sender's utility. In the canonical Bayesian persuasion model, the Sender's utility at every belief is well defined by the primitives (see the "$\hat{v}$" function in KG). However, in our model, the Sender's utility at each belief includes both his payoff *and* his messaging cost, but the latter is an equilibrium object. The second complication arises from the additional stage-2 communication game here, which means that some care is required to ensure that the chosen combination of ($|\Omega|$ or less) beliefs and utilities can be sustained as an equilibrium in the communication game.

5 Equilibrium

One of KG's methodological insights is that the canonical Bayesian persuasion problem reduces to finding a distribution of beliefs that averages back to the prior, which KG call "Bayes plausibility". With an additional stage-2 communication game, our model might not be amenable to this belief-based approach because the Sender's messaging strategy must be determined in equilibrium, and the players' equilibrium beliefs can also differ. The importance of Proposition 1 is that when Condition 1 or 2 holds, the Sender-optimal equilibrium outcome can always be supported by the two players holding the same ex-post belief. Therefore, we have to characterize only one set of belief distributions. Moreover, from Proposition 2, we can restrict our attention to distributions supported on at most $|\Omega|$ different beliefs. The Sender's value from persuasion then reduces to optimizing over the set of Bayes plau-

[14] The literature typically appeals to the Carathéodory theorem to show that only a finite number of signals are required. The theorem states that any element x in the convex hull of a set $X \subset \mathbb{R}^n$ can be written as a convex combination of $n + 1$ or fewer elements in X. The Fenchel-Bunt theorem, which is used here, refines the Carathéodory theorem, and it states that x can be written as n or less elements in X if X is also a connected set. See, for example, Hiriart-Urruty and Lemaréchal (2012), Theorem 1.3.6 (pp. 29) and Theorem 1.3.7 (pp.30).

sible distributions of beliefs with a cost attached to each distribution. We illustrate this in Section 5.1 below and then illustrate in Section 5.2 that the Sender might still achieve his full commitment payoff in the absence of full commitment power.

5.1 From costly messages to costly beliefs

Since we are concerned with just one set of beliefs now, we switch notation slightly and use β to denote a belief in $\Delta\Omega$ from now on. Let $\mathcal{D} \subset \Delta\Delta\Omega$ be the set of Bayes plausible distributions of belief supported on $|\Omega|$ or less beliefs. We represent each $\tau \in \mathcal{D}$ by two $1 \times |\Omega|$ vectors: $\tau = \left\{\vec{\beta}; \vec{\delta}\right\} = \left\{(\beta_j)_{j=1,\ldots,|\Omega|}; (\delta_j)_{j=1,\ldots,|\Omega|}\right\}$, where $\beta_j \in \Delta\Omega\ \forall j$ and $\vec{\delta}$ is a probability vector such that $\sum_{j=1}^{|\Omega|} \delta_j \beta_j = \beta^o$. Given a Receiver's action strategy α, the Sender's payoff from belief distribution $\tau = \left\{\vec{\beta}; \vec{\delta}\right\} \in \mathcal{D}$ is

$$V^* \left(\left\{\vec{\beta}; \vec{\delta}\right\}, \alpha\right) = \sum_{j=1}^{|\Omega|} \delta_j v\left(\alpha\left(\cdot|\beta_j\right), \beta_j\right).$$

Let

$$C^* \left(\left\{\vec{\beta}; \vec{\delta}\right\}, \alpha\right) := \min_{\substack{s_1,\ldots,s_{|\Omega|}\in S \\ m_1,\ldots,m_{|\Omega|}\in M}} \sum_{j=1}^{|\Omega|} \delta_j c\left(m_j|s_j\right) \tag{6}$$

subject to

$$v\Big(\alpha\left(\cdot|\beta_j\right), \beta_j\Big) - c\left(m_j|s_j\right) \geq v\Big(\alpha\left(\cdot|\beta_{j'}\right), \beta_j\Big) - c\left(m_{j'}|s_j\right) \quad \forall j, j' = 1, \ldots, |\Omega|, \tag{7}$$

with the convention that $C^* \left(\left\{\vec{\beta}; \vec{\delta}\right\}, \alpha\right) = \infty$ when the feasible set in (7) is empty. Define program $\mathcal{P}$ as:

$$\max_{\tau\in\mathcal{D},\alpha\in\mathcal{A}^*} V^*\left(\tau,\alpha\right) - C^*\left(\tau,\alpha\right) \tag{$\mathcal{P}$}$$

Theorem 1. *Suppose that Condition 1 or 2 holds. A Sender-optimal equilibrium exists if and only if a solution to program $\mathcal{P}$ exists. Moreover, if W^* is the value for program $\mathcal{P}$, then W^* is the Sender's value from persuasion.*

As in the canonical Bayesian persuasion model, each belief distribution is associated with a Sender's expected payoff. The difference here is the additional Sender's expected messaging cost C^*, which can be interpreted as the cost of generating each belief distribution. In contrast to papers using information measures such as Shannon's (1948) entropy (or a generalized version of it) to determine the cost of a signal structure,[15] C^* is a cost to sustain

[15] E.g., Gentzkow and Kamenica (2014); Matyskova (2018).

the resulting belief distribution as an equilibrium in the communication game due to the Sender's lack of commitment to truthfully revealing the realized signal. In particular, the cost is infinite if the feasible set in (7) is empty; this means that any information that cannot be credibly transmitted under the institution cannot be generated in the first place.

Since any Bayes plausible distribution of beliefs can be generated by some signal structure (see KG), the key to establishing Theorem 1 concerns the set of Bayes plausible beliefs sustainable in equilibrium. From Propositions 1 and 2, this will be incorporated by the set of constraints in (7). Theorem 1 thus implies that the Sender's value from persuasion can be solved by a two-step process: first, we derive the cost function C^* for any given belief distribution for the Receiver; then, we maximize over the belief distribution.

5.2 Full commitment payoff without full commitment

Theorem 1 implies that if an institution can allow for any $|\Omega|$ beliefs to always be credibly transmitted, the Sender can achieve his full commitment payoff regardless of the players' preferences. The following proposition provides a few classes of institutions with this property.

Proposition 3. *The Sender's value from persuasion is always his full commitment payoff under the following institutions:*

- *Quadratic lying cost with* $\mathcal{S}^B = \mathbb{R}$.
- *"Distance" lying cost with* $\mathcal{S}^B = \mathbb{R}$ *and* $d(x, y) = |x - y|$.
- *Partial verifiability with* $|\mathcal{V}| \geq |\Omega|$, *or with* $|\mathcal{V}| = |\Omega| - 1$ *and the Sender having state-independent preferences.*

The first parts of Proposition 3 follow from noting that with a quadratic lying cost (and analogously, a "linear distance" lying cost), if the set of available signals is sufficiently large, there will exist a set of signals that is sufficiently "differentiated" in the sense that lying across signals within this set is too costly for the Sender. In turn, using this set of signals maintains the Sender's credibility costlessly and allows him to achieve his full commitment payoff. As for the institution with partial verifiability, when $|\mathcal{V}| \geq |\Omega|$, there are sufficient verifiable signals to provide the credibility for the $|\Omega|$ beliefs needed to attain the Sender's full commitment payoff. If the Sender also has state-independent preferences, the beliefs can be ranked. At the Sender's most preferred belief, his credibility to not convey other inferior beliefs is undoubted, so only $|\Omega| - 1$ pieces of verifiable signals are needed to convey these other beliefs.

Although the arguments behind Proposition 3 are simple, the result is less straightforward *a priori*. This is because most of these institutions do not have an intuitive sense of providing the Sender with full commitment to truthfully revealing information. Moreover, truth-telling is also never an equilibrium in the associated communication games in isolation. In particular, Kartik et al. (2007) show that in a strategic communication game with a quadratic lying cost, there is full information transmission in equilibrium, but it is always via a language with costly inflation; Dye (1985) and Jung and Kwon (1988) show that under a verifiable disclosure communication game, if the Sender can feign ignorance, there is not even full information transmission.

6 Constant Lying Cost and Partial Commitment

In this section, we focus on the institution with a constant lying cost, which allows us to quantify the Sender's commitment level as the value of k. We first note a general feature of such institutions:

Lemma 2. *Under the institution with a constant lying cost k, for any $\alpha \in \mathcal{A}^*$ and $\tau = \left\{ (\beta_j)_{j=1,\ldots,|\Omega|} ; (\delta_j)_{j=1,\ldots,|\Omega|} \right\} \in \mathcal{D}$,*

$$C^* (\tau, \alpha) = \begin{cases} 0 & \textit{if the no-lying constraint is satifised,} \\ \infty & \textit{otherwise,} \end{cases}$$

where the no-lying constraint is satisfied if

$$v\Big(\alpha (\cdot|\beta_{j'}), \beta_j\Big) - v\Big(\alpha (\cdot|\beta_j), \beta_j\Big) \leq k \qquad \forall j, j' = 1, \ldots, |\Omega| . \tag{8}$$

There is no lying in *any* Sender-optimal equilibrium here, but the no-lying condition puts a constraint on the set of sustainable belief distributions. The set, in turn, expands (in the set inclusion sense) with k, so the Sender's value from persuasion weakly increases in the Sender's commitment level k. We consider a few types of Sender's preferences under this institution next.

6.1 Quadratic loss payoffs

Let Ω be a finite subset of $[0,1]$, and $A = M = \mathcal{S}^{\mathcal{B}} = [0,1]$. Let the Receiver's and the Sender's payoffs be, respectively,

$$u(a,\omega) = -(a-\omega)^2, \tag{9}$$

$$v(a,\omega) = -(a-\omega-b)^2, \tag{10}$$

where $b > 0$. The quadratic loss payoffs capture the notion that actions are progressively less preferred by the Receiver the further they are away from the true state at either side, and b quantifies a constant bias that the Sender has for the ideal action. This is a common specification in the communication literature and has been applied widely in areas such as political economy and organizational economics. When endowed with a uniform prior, this is also called the "uniform-quadratic model". Our aim here is to characterize a prior-free sufficient commitment level for the Sender that allows him to achieve his full commitment payoff.

Under any Sender's belief $\beta \in \Delta\Omega$ and Receiver's action a, the Sender's expected payoff is

$$E_\beta\left[-(a-\omega-b)^2\right] = -Var_\beta(\omega) - (a-b-E_\beta(\omega))^2. \tag{11}$$

It is straightforward to show that when the Receiver holds belief $\beta \in \Delta\Omega$, her optimal action is uniquely $a = E_\beta(\omega)$. Therefore, the no-lying constraint in (8) here is

$$\left(E_{\beta_{j'}}(\omega) - E_{\beta_j}(\omega) - b\right)^2 \geq b^2 - k \quad \forall j, j' = 1, \ldots, |\Omega|. \tag{12}$$

Proposition 4. *When the players' preferences are represented by the quadratic loss functions in (9) and (10) and the information transmission takes place under an institution with a constant lying cost k, the Sender can obtain his full commitment payoff whenever $k \geq b^2$.*

This implies that the required commitment to achieve the Sender's full commitment payoff is $k = b^2$. This is because (12) is always satisfied when $k \geq b^2$, which implies that $C^*(\tau, \alpha) = 0$ for any $\tau \in \mathcal{D}$ and $\alpha \in \mathcal{A}^*$. Proposition 4 thus follows immediately from Theorem 1. The Sender's payoff (hence utility) in (11) is maximized by choosing the fully informative signal structure so that the variance is zero, and he obtains his full commitment payoff of $-b^2$.

6.2 State-independent preferences for the Sender

The "concavification" method (Aumann and Maschler, 1995; KG) provides a geometrical characterization of the optimal signal structure under full commitment, but it is generally not amenable to our setup due to the additional messaging cost. However, since the constant lying cost simply puts a constraint on the set of feasible signal structures, the problem can admit a similar geometrical characterization if the constraint can also be represented geometrically. This will be the case when the Sender has state-independent preferences.

Suppose now that the Sender's payoff function satisfies $v(a,\omega) = \nu(a)\ \forall a \in A, \omega \in \Omega$ (i.e., the Sender's payoff is independent of the state). Define the correspondence $\hat{v} : \Delta\Omega \rightrightarrows A$, where for all $\beta \in \Delta\Omega$,

$$\hat{v}(\beta) = \left\{ co(\nu(a)) \,|a \in \bar{A}(\beta) \right\}, \tag{13}$$

where "*co*" denotes the convex hull.[16]

Figure 1: State-independent preferences under constant lying cost and $|\Omega| = 2$.

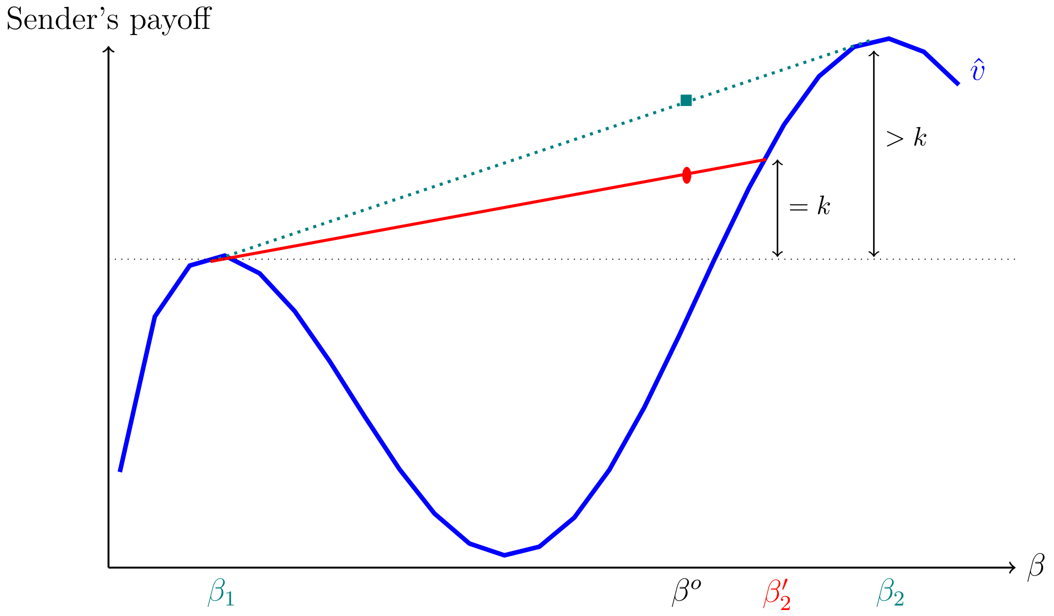


Figure 1 plots the $\hat{v}$ graph for $|\Omega| = 2$. A signal structure is represented by a line joining any two points on $\hat{v}$, and the no-lying constraint in (8) restricts the vertical distance between

[16] We note that our definition of $\hat{v}$ is slightly different from that in KG. As noted earlier, in KG, under any belief, the Receiver is assumed to break indifference in favor of the Sender, so $\hat{v}$ in KG is a *function* that gives the Sender's (indirect) payoff under each Receiver's belief. In contrast, $\hat{v}$ here is a *correspondence* that gives the set of Sender's (indirect) payoffs in which the Receiver is optimizing given her belief.

these two points on $\hat{v}$ to be less than k. The green dotted line is the concave closure of $\hat{v}$,[17] and the Sender's full commitment payoff is the value of the concave closure at the prior β^o, which is marked by the shaded square. The optimal signal structure under full commitment generates beliefs β_1 and β_2, but the no-lying constraint is violated for this distribution of beliefs. Instead, the Sender must induce a less valuable distribution of beliefs – namely, β_1 and β_2' – which gives him a value of persuasion marked by the red dot.

6.3 An example of the Receiver's wariness

In this subsection, we use an example to illustrate some notions of the Receiver's wariness that are absent when the Sender has full commitment. Let $\Omega = \{0, 1\}$ and $A = [0, 1]$. The Sender has a state-independent payoff function $\nu(a)$, which is strictly increasing and convex,[18] with $\nu(0) = 0$. The Receiver receives a constant marginal benefit of 1 (which is a normalization) from her action when $\omega = 1$, and she incurs a constant marginal loss of $L > 0$ from her action when $\omega = 0$. In addition, action is costly and she incurs a quadratic cost of $\frac{a^2}{2}$ for taking action a regardless of the state. Jointly, the Receiver's payoff function can be written as $u(a, \omega) = [\omega - L(1 - \omega)] a - \frac{a^2}{2}$.

Such a setup is applicable to many economic situations. For example, the Receiver could be a politician whose policy choice is parametrized by a, where a smaller a represents more left-wing policies and a larger a represents more right-wing policies. The politician is left-leaning and hence incurs a private cost for implementing right-wing policies, but she is also concerned about her re-election and has to pander her policies to her voters' preferences. Let $\omega = 0$ denote a more liberal group of voters (who prefer small a) and $\omega = 1$ denote a more conservative group of voters (who prefer big a). The politician is uncertain about ω, and the Sender represents a right-leaning lobbyist generating information through, say, voter research to persuade the politician to implement more right-wing policies. Away from political economy, the setup is also applicable to marketing and advertising – the Sender is a seller who is generating information about an unknown state of the world that determines the buyer's (Receiver) utility from the seller's good, such as insurance and other financial products.

In the situations just described, the Sender is often able to misrepresent the new information at some potential cost. The lobbyist could commit to the type of voter research

[17] The concave closure of $\hat{v}$ is a function $\hat{V}(\beta) := \{\sup\ z \mid (\beta, z) \in co(\hat{v})\}$.

[18] The results also hold if ν is not too concave for any given set of parameter values of L and k defined below.

but then lie about its result at a cost derived from, for example, the "effort" to doctor the statistics and the potential backlashes when being caught doing so. Similarly, the seller could commit to the type of research but lie about its outcome at a cost that is similar to that above. We account for this by modeling the information gathering and transmission process as taking place under an institution with a constant lying cost. From Proposition 2, two different signals will be sufficient for our binary state space, so we let $\mathcal{S}^B = M = \{0, 1\}$, and the Sender faces a constant lying cost $k > 0$.

To simplify notation, we denote belief β as the probability of $\omega = 1$. It is readily noted that the Receiver's optimal action is uniquely $\max\{0, \beta - L(1-\beta)\}$ when she holds belief β. Therefore,

$$\hat{v}(\beta) = \begin{cases} \left\{\nu(0)\right\} & \text{, if } \ \beta \leq \frac{L}{1+L} \\ \left\{\nu(\beta - L(1-\beta))\right\} & \text{, if } \ \beta > \frac{L}{1+L} \end{cases}$$

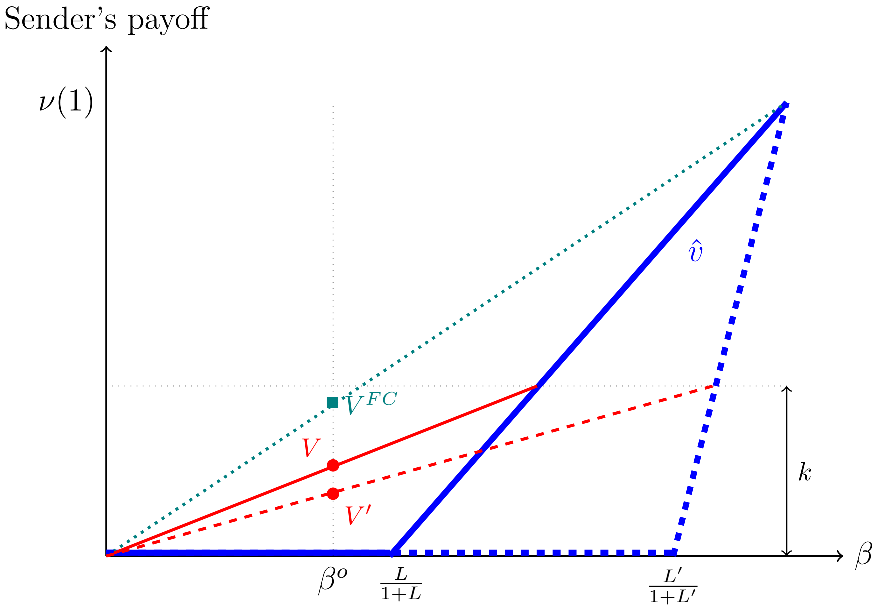
We assume that $\beta^o \leq \frac{L}{1+L}$, so the Sender gets his worst payoff without further information to the Receiver. It is readily verified that the concave closure of $\hat{v}$ is $\nu(1)\beta$, which is the straight line joining $\nu(0)$ and $\nu(1)$. Therefore, with full commitment, the concavification method implies that the optimal signal structure under any prior is the fully informative one. This implies that the Sender achieves his full commitment payoff when $k \geq \nu(1)$ because the no-lying constraint (8) for the fully informative signal structure is satisfied. Therefore, we will only consider $k < \nu(1)$.

Proposition 5. *Suppose that $k < \nu(1)$. The Sender's value from persuasion is $\frac{k\beta^o(1+L)}{\nu^{-1}(k)+L}$, and the optimal signal structure generates beliefs 0 and $\bar{\beta} = \frac{\nu^{-1}(k)+L}{1+L} < 1$ with respective probabilities $1 - \frac{\beta^o(1+L)}{\nu^{-1}(k)+L}$ and $\frac{\beta^o(1+L)}{\nu^{-1}(k)+L}$. $\bar{\beta}$ increases with both k and L; the Sender's value from persuasion increases with k but decreases with L.*

Figure 2 illustrates the case for a linear ν. The blue solid graph is $\hat{v}$ and the green dotted line plots its concave closure. Therefore, the Sender's full commitment payoff is the value on the concave closure at the prior, which is marked by V^{FC}. The optimal signal structure is represented by the red solid line, which joins the origin to the $\hat{v}$ graph at a height of k; the Sender's value from persuasion is the value on this line at the prior, which is marked by V. It is readily observed that a higher k would give a steeper line, which, in turn, gives a higher Sender's value from persuasion.

When L increases to L', the $\hat{v}$ graph shifts to the dashed blue graph. The concave closure of $\hat{v}$ remains unchanged, so the Sender's full commitment payoff remains the same as well. However, the line representing the optimal signal structure is now lower, as represented

Figure 2: Receiver's wariness under $A = [0,1]$ and linear ν.



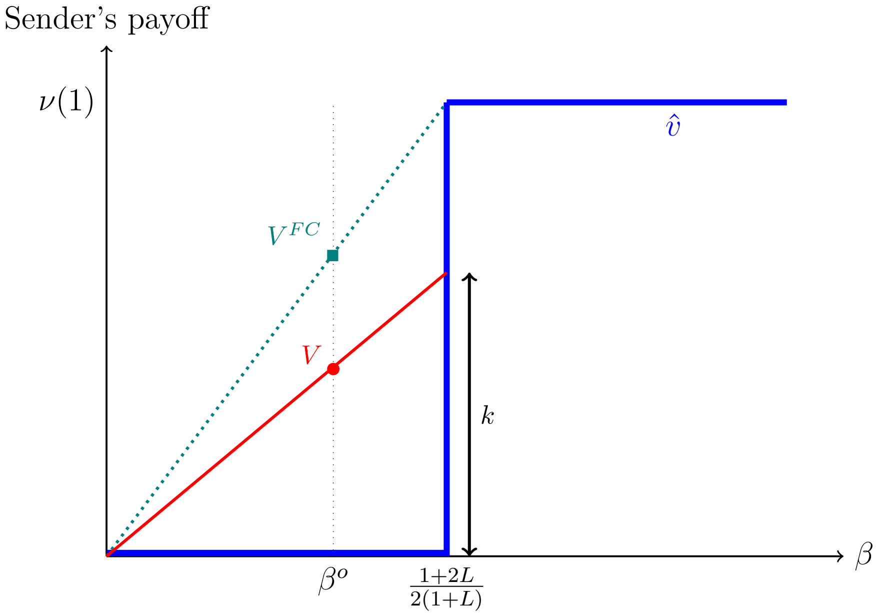
by the dashed red line. The new optimal signal structure generates a larger spread in the Receiver's posteriors, but the Sender's value from persuasion drops to V'.

Proposition 5 captures two notions of the Receiver's wariness that make persuasion more "difficult" for the Sender. The first is straightforward. When the Sender's lying cost decreases (i.e., lower k), the Sender's ability to generate a wider spread in the Receiver's posterior beliefs diminishes. This is because a wider spread in the Receiver's beliefs implies a greater difference in the Sender's payoffs across beliefs, which increases the Sender's gain from lying. Since the Sender benefits from spreading the Receiver's beliefs here, the inability to do so decreases the Sender's value from persuasion.

The second notion of wariness arises in the Receiver's response when the cost of making mistakes about the state becomes higher, as represented by an increase in L. Under any interior belief, the Receiver will take a lower action when the cost of making mistakes becomes higher. In the absence of full commitment, the Sender cannot credibly let the Receiver learn about the Sender's preferred state (i.e., $\omega = 1$), so the Receiver's conservativeness against a higher action adversely affects the Sender's value from persuasion. By contrast, this effect has no bite when the Sender has full commitment because the Sender will always let the Receiver learn the state fully; therefore, a change in the value of L has no effect on the Sender's full commitment payoff.

The Receiver exhibits a third form of wariness when her action set is binary. If $A = \{0, 1\}$,

Figure 3: Receiver's wariness under $A = \{0, 1\}$.



the Receiver will adopt a threshold action strategy with respect to her belief, which generates the $\hat{v}$ correspondence represented by the blue solid graph in Figure 3 – the Receiver chooses $a = 0$ $(a = 1)$ under belief $\beta < (>) \frac{1+2L}{2(1+L)}$, and she is indifferent between either action when $\beta = \frac{1+2L}{2(1+L)}$. With full commitment and the Receiver taking the Sender's preferred action of $a = 1$ whenever she is indifferent (i.e., the KG assumptions), the optimal signal structure generates beliefs 0 and $\frac{1+2L}{2(1+L)}$, and the Sender's full commitment payoff is marked by V^{FC}. However, when the Sender can lie about the signal at a cost $k < \nu(1)$, the Sender will always report the signal that generates the higher belief if the Receiver were to always choose $a = 1$ after it. Therefore, this cannot be an equilibrium. Nevertheless, the optimal signal structure in the absence of full commitment remains unchanged, but the Receiver must exhibit wariness by not always choosing the Sender's preferred action when she holds belief $\frac{1+2L}{2(1+L)}$. In particular, she plays a mixed strategy at belief $\frac{1+2L}{2(1+L)}$ and chooses $a = 1$ with a probability of $\frac{k}{\nu(1)} < 1$, which then removes the Sender's lying incentive. This results in a lower Sender's value from persuasion, which is marked by V, and it is readily verified that $V = \frac{k}{\nu(1)} V^{FC}$. This example also illustrates why the Sender-optimal equilibrium does not always entail the Receiver taking the Sender-preferred action when she is indifferent (see Remark 1).

7 Belief-dependent Communication Costs

Our analysis thus far has focused solely on the communication costs related to "encoding" the information within the signal space and then communicating it within the message space of the institution. What we have not considered is a penalty on the Sender for misrepresenting actual information, which are the beliefs over the states. This is without much loss in equilibrium, since our results rest on equilibria exhibiting full information revelation (equivalently, no information misrepresentation). However, there might be a concern that the presence of information misrepresentation costs might disrupt such equilibria.

In this section, we impose an additional cost on the Sender for information misrepresentation. To present this extension, we return to the convention of using "σ" to denote the Sender's belief and "ρ" to denote the Receiver's belief. Let the sequence of the game be unchanged from the baseline model, and the Receiver's utility be as described there. The Sender's utility is now

$$v\left(a,\omega\right)-c\left(m|s\right)-\psi\left(\rho|\sigma\right),$$

where v and c are, respectively, his payoff and messaging cost as before. The additional $\psi : \Delta\Omega \times \Delta\Omega \to \mathbb{R}^+$ is a non-negative function, where $\psi\left(\rho|\sigma\right)$ is the Sender's cost of inducing the Receiver to have belief ρ when he holds belief σ.[19]

The strategies of both players are unchanged from Section 3.2. The definition of a PBE in the stage-2 communication game remains that as defined in Definition 1, except with a modification on the Sender's best response in (S-IC) to the following: for any $s \in S$ and $m \in M$ in the support of $\mu\left(\cdot|s\right)$,

$$\begin{aligned}&v\Big(\alpha\left(\cdot|\rho_{\pi,\mu}\left(\cdot|m\right)\right),\sigma_{\pi}\left(\cdot|s\right)\Big)-c\left(m|s\right)-\psi\Big(\rho_{\pi,\mu}\left(\cdot|m\right)\Big|\sigma_{\pi}\left(\cdot|s\right)\Big)\\ \geq\ &v\Big(\alpha\left(\cdot|\rho_{\pi,\mu}\left(\cdot|m'\right)\right),\sigma_{\pi}\left(\cdot|s\right)\Big)-c\left(m'|s\right)-\psi\Big(\rho_{\pi,\mu}\left(\cdot|m'\right)\Big|\sigma_{\pi}\left(\cdot|s\right)\Big)\quad\forall m'\in M.\end{aligned}\tag{S-IC'}$$

As in (5), we let $W\left(\pi,\mu;\alpha\right)=V\left(\pi,\mu;\alpha\right)-C\left(\pi,\mu\right)$, where V is unchanged from (3), whereas C is now

$$C\left(\pi,\mu\right):=\sum_{\omega\in\Omega}\beta^{o}\left(\omega\right)\int_{s\in S}\int_{m\in M}\Big[c\left(m|s\right)+\psi\Big(\rho_{\pi,\mu}\left(\cdot|m\right)\Big|\sigma_{\pi}\left(\cdot|s\right)\Big)\Big]d\mu\left(m|s\right)d\pi\left(s|\omega\right).$$

[19] Note that to ease exposition, we have simplified notation significantly for ψ. In particular, ρ is an equilibrium object derived from Bayes rule; thus, it depends on the signal structure and the Receiver's conjecture about the Sender's messaging strategy.

Assumption 2. *The function ψ satisfies:*

1. *for any given $\sigma, \rho \in \Delta\Omega$, $\psi(\rho|\sigma) = 0$ if $\rho = \sigma$;*
2. *for any given $\rho, \rho' \in \Delta\Omega$ and $\tau \in \Delta\Delta\Omega$ such that $\int \sigma d\tau(\sigma) = \rho$,*

$$\psi(\rho'|\rho) \geq \int \psi(\rho'|\sigma) - \psi(\rho|\sigma) d\tau(\sigma).$$

Assumption 2.1 is a natural assumption that the cost of no information misrepresentation is always zero. Assumption 2.2 is a technical condition that, by Lemma 3 below, is satisfied by many commonly used divergence measures.

Lemma 3. *Assumption 2 is satisfied by the following divergence measures:*

- *Euclidean distance: $\psi(\rho|\sigma) = \sqrt{\sum_{\omega\in\Omega}(\sigma(\omega) - \rho(\omega))^2}$.*
- *Squared Euclidean distance: $\psi(\rho|\sigma) = \sum_{\omega\in\Omega}(\sigma(\omega) - \rho(\omega))^2$.*
- *Kullback-Leiber divergence: $\psi(\rho|\sigma) = \sum_{\omega\in\Omega}\sigma(\omega)\log\frac{\sigma(\omega)}{\rho(\omega)}$.*
- *Jeffreys divergence: $\psi(\rho|\sigma) = \sum_{\omega\in\Omega}(\sigma(\omega) - \rho(\omega))\log\frac{\sigma(\omega)}{\rho(\omega)}$.*[20]

Proposition 6. *Suppose that the Sender's utility has an additional belief-dependent communication cost ψ that satisfies Assumption 2, and a Sender-optimal equilibrium exists. If Condition 2 holds, there exists a Sender-optimal equilibrium with full information revelation (cf. Proposition 1), and the numbers of signals and messages used in equilibrium are, respectively, less than or equal to $|\Omega|$ (cf. Proposition 2).*

When there is full information revelation, the two players' posteriors always coincide, which implies that the Sender's equilibrium belief-dependent communication cost ψ is always zero. Therefore, all our analysis following from Propositions 1 and 2 follows through. Moreover, Proposition 1 is actually strengthened by the additional cost ψ in the following sense: if $\psi(\rho|\sigma) > 0$ whenever $\rho \neq \sigma$, then under Condition 2, *all* Sender-optimal equilibria must exhibit full information revelation.

[20] The Jeffreys divergence is first introduced in Jeffreys (1946). It is also called the J-divergence or the symmetric divergence.

8 Conclusion

In this paper, we have provided a model of Bayesian persuasion where the Sender does not have full commitment to reporting all new information to the Receiver. We define an institution, which is distinct from the players' preferences over actions, as the triple of the signal space for generating new information, the message space for communicating the new information to the Receiver, and the associated messaging cost at each signal realization. The Sender's commitment is thus derived from the institution. We provide sufficient conditions under which there is full revelation of new information to the Receiver. This reduces the problem to the familiar belief-based approach of looking for the optimal distribution of Receiver's beliefs, with the additional factor of a cost to sustain the belief distribution as an equilibrium in the communication subgame. Using this, we show that some institutions continue to allow the Sender to achieve his full commitment persuasion payoff even though the associated communication games in isolation never admit a truthful-reporting equilibrium. The particular class of institution with a constant lying cost provides a way to quantify the Sender's commitment level and study its effect on the Sender's value from persuasion, thus helping to bridge the cheap talk communication literature and the information design literature.

References

Argenziano, R., S. Severinov, and F. Squintani (2016). Strategic information acquisition and transmission. *American Economic Journal: Microeconomics 8*(3), 119–55.

Aumann, R. J. and M. Maschler (1995). *Repeated Games with Incomplete Information*. MIT press.

Austen-Smith, D. and J. S. Banks (2000). Cheap talk and burned money. *Journal of Economic Theory 91*(1), 1–16.

Best, J. and D. Quigley (2017). Persuasion for the long run.

Bester, H. and R. Strausz (2001). Contracting with imperfect commitment and the revelation principle: the single agent case. *Econometrica 69*(4), 1077–1098.

Crawford, V. P. and J. Sobel (1982). Strategic information transmission. *Econometrica 50*(6), 1431–1451.

Deb, R. and M. Said (2015). Dynamic screening with limited commitment. *Journal of Economic Theory 159*, 891–928.

Doval, L. and V. Skreta (2018). Mechanism design with limited commitment.

Dye, R. A. (1985). Disclosure of nonproprietary information. *Journal of Accounting Research*, 123–145.

Fréchette, G., A. Lizzeri, and J. Perego (2018). Rules and commitment in communication.

Fuchs, W. (2015). Subjective evaluations: Discretionary bonuses and feedback credibility. *American Economic Journal: Microeconomics 7*(1), 99–108.

Gentzkow, M. and E. Kamenica (2014). Costly persuasion. *American Economic Review 104*(5), 457–62.

Gentzkow, M. and E. Kamenica (2017). Disclosure of endogenous information. *Economic Theory Bulletin 5*(1), 47–56.

Grossman, S. J. (1981). The informational role of warranties and private disclosure about product quality. *The Journal of Law and Economics 24*(3), 461–483.

Guo, Y. and E. Shmaya (2018). Costly miscalibration in communication.

Hiriart-Urruty, J.-B. and C. Lemaréchal (2012). *Fundamentals of Convex Analysis*. Springer Science & Business Media.

In, Y. and J. Wright (2017). Signaling private choices. *The Review of Economic Studies 85*(1), 558–580.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 453–461.

Jung, W.-O. and Y. K. Kwon (1988). Disclosure when the market is unsure of information endowment of managers. *Journal of Accounting Research*, 146–153.

Kamenica, E. (Forthcoming). Bayesian persuasion and information design. *Annual Review of Economics*.

Kamenica, E. and M. Gentzkow (2011). Bayesian persuasion. *American Economic Review 101*(6), 2590–2615.

Karamychev, V. and B. Visser (2017). Optimal signaling with cheap talk and money burning. *International Journal of Game Theory 46*(3), 813–850.

Kartik, N. (2007). A note on cheap talk and burned money. *Journal of Economic Theory 136*(1), 749–758.

Kartik, N. (2009). Strategic communication with lying costs. *The Review of Economic Studies 76*(4), 1359–1395.

Kartik, N., M. Ottaviani, and F. Squintani (2007). Credulity, lies, and costly talk. *Journal of Economic theory 134*(1), 93–116.

Kolotilin, A. and H. Li (2018). Relational communication.

Lipnowski, E. and D. Ravid (2017). Cheap talk with transparent motives.

Lipnowski, E., D. Ravid, and D. Shishkin (2018). Persuasion via weak institutions.

Matyskova, L. (2018). Bayesian persuasion with costly information acquisition.

Milgrom, P. R. (1981). Good news and bad news: Representation theorems and applications. *The Bell Journal of Economics*, 380–391.

Min, D. (2017). Bayesian persuasion under partial commitment.

Pei, H. (2015). Communication with endogenous information acquisition. *Journal of Economic Theory 160*, 132–149.

Rayo, L. and I. Segal (2010). Optimal information disclosure. *Journal of Political Economy 118*(5), 949–987.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal 27*(3), 379–423.

Shiryaev, A. N. (1996). *Probability.* New York: Springer.

Skreta, V. (2006). Sequentially optimal mechanisms. *The Review of Economic Studies 73*(4), 1085–1111.

Skreta, V. (2015). Optimal auction design under non-commitment. *Journal of Economic Theory 159*, 854–890.

Spence, M. (1973). Job market signaling. *The Quarterly Journal of Economics 87*(3), 355–374.

Verrecchia, R. E. (1983). Discretionary disclosure. *Journal of Accounting and Economics 5*, 179–194.

A Proof Appendix

Proof of Lemma 1.

Fix a $m \in M$ and a $\Lambda \in \Delta S$ with the property that $m \in \mathcal{M}(\Lambda)$.

"Distance" lying cost. Set $\bar{s}$ to be a signal with a base signal $m \in \mathcal{S}^B$. For the first property, the LHS is $d(m,m) = 0$ and the RHS is weakly positive, so the first property holds. Next, since d is a metric, it satisfies the triangle inequality: $d(m', s) \leq d(m', m) + d(m, s)$. Therefore, $\forall m' \in \mathcal{M}(\Lambda)$,

$$\begin{aligned} c(m|\bar{s}) - c(m'|\bar{s}) =& 0 - \int_{s \in S} d(m', m)\, d\Lambda(s) \\ &\leq \int_{s \in S} d(m,s) - d(m', s)\, d\Lambda(s) \\ &= \int_{s \in S} c(m|s) - c(m'|s)\, d\Lambda(s), \end{aligned}$$

which implies the second property. For the third property, for any $m' \neq m$,

$$\begin{aligned} \int_{s \in S} c(m'|s)\, d\Lambda(s) &= \int_{s \in S} d(m', s)\, d\Lambda(s) \\ &\leq \int_{s \in S} d(m,s) + d(m', m)\, d\Lambda(s) \\ &= \int_{s \in S} c(m|s)\, d\Lambda(s) + d(m', m). \end{aligned}$$

Since $m \in \mathcal{M}(\Lambda)$, $\int_{s \in S} c(m'|s)\, d\Lambda(s) < \infty$. This implies that $m' \in \mathcal{M}(\Lambda)$ $\forall m'$; therefore, the third property is trivially satisfied.

Constant lying cost. The constant lying cost can be represented by $c(m|s) = d(m,s)$, where $d(x,y) = 0$ if $x = y$ and $d(x,y) = k$ if $x \neq y$. Notice that for any x, y, z, if $x = y = z$, $d(x,y) + d(y,z) = 0 = d(x,z)$; if $x = y \neq z$, $d(x,y) + d(y,z) = k = d(x,z)$; if $x \neq y \neq z$ and $x \neq z$, $d(x,y) + d(y,z) = 2k > k = d(x,z)$. Therefore, d satisfies the triangle inequality: $d(x,y) + d(y,z) \geq d(x,z)$ $\forall x, y, z$.[21] The previous argument for the "Distance" lying cost thus follows through here.

[21] The constant lying cost is essentially the "discrete metric" except that the discrete metric is typically defined as $d(x,y) = 1$ (instead of k) for any $x \neq y$.

Quadratic lying cost. Set $\bar{s} = \int s d\Lambda(s)$. Since S is a convex set, $\bar{s} \in S$. $(m-s)^2$ is convex in s for any m; therefore, the first property holds. Next, for any m, m', $(m-s)^2 - (m'-s)^2 = m^2 - (m')^2 - 2(m-m')s$, which is linear in s; therefore, the second property holds. For the third property, first note that for any $x, y \in \mathbb{R}$,

$$(x+y)^2 = 2x^2 + 2y^2 - (x-y)^2 \leq 2x^2 + 2y^2. \tag{14}$$

For any $m' \neq m$,

$$\begin{aligned} \int_{s\in S} c(m'|s)\, d\Lambda(s) =& \int_{s\in S} (s-m+m-m')^2\, d\Lambda(s) \\ \leq& \int_{s\in S} 2(s-m)^2 + 2(m-m')^2\, d\Lambda(s) \\ =& \, 2\int_{s\in S} c(m|s)\, d\Lambda(s) + 2(m-m')^2, \end{aligned} \tag{15}$$

where the inequality in (15) follows from (14). Since $m \in \mathcal{M}(\Lambda)$, $\int_{s\in S} c(m'|s)\, d\Lambda(s) < \infty$. This implies that $m' \in \mathcal{M}(\Lambda)\ \forall m'$; therefore, the third property is trivially satisfied.

Partial Verifiability. If $m \neq \phi$, set $\bar{s}$ to be a signal with a base signal $m \in \mathcal{S}^B$; if $m = \phi$, set $\bar{s}$ to be any signal with a base in $\mathcal{N}$. Therefore, $c(m|\bar{s}) = 0$, and the first property holds. For the second property, consider a $m' \in \mathcal{M}(\Lambda)$, which implies that $\int_{s\in S} c(m'|s)\, d\Lambda(s) = 0$; therefore, the RHS is weakly positive. Since $m \in \mathcal{M}(\Lambda)$, it must hold that $\int_{s\in S} c(m|s)\, d\Lambda(s) = 0$, so the LHS is 0, and the second property thus holds. For the third property, if $m' \in M \backslash \mathcal{M}(\Lambda)$, then $m' \in \mathcal{V}$, which implies that $c(m'|\bar{s}) = \infty$.

Costly disclosure. If $m \neq \phi$, set $\bar{s}$ to be a signal with a base signal $m \in \mathcal{S}^B$; if $m = \phi$, set $\bar{s}$ to be any signal with a base in $\mathcal{N}$. If $m \in \mathcal{V}$, then the values of both $c(m|\bar{s})$ and $\int_{s\in S} c(m|s)\, d\Lambda(s)$ must be k;[22] if $m \in \mathcal{N}$ or $m = \phi$, then the values of both must be 0. Therefore, the first property holds, and the LHS of the second property is always zero. Consider a $m' \in \mathcal{M}(\Lambda)$. If $m' \in \mathcal{N}$ or $m' = \phi$, then the RHS is also zero; if $m' \in \mathcal{V}$, then $\int_{s\in S} c(m'|s)\, d\Lambda(s) = k$ while $c(m'|\bar{s}) = \infty$ because $m' \neq m$. Therefore, the second property is also satisfied. Next, consider a $m' \notin \mathcal{M}(\Lambda)$. This must imply that $m' \in \mathcal{V}$, so $c(m'|\bar{s}) = \infty$ because $m' \neq m$, which implies the third property.

[22] In particular, all signals supported on Λ must have base signal m.

Proof of Proposition 1.

Proof. Let $(\mu;\alpha)$ be a PBE of the stage-2 communication game under π. Let $\tilde{S}\subseteq S$ be the set of signals in the support of $\{\pi(\cdot|\omega)\}_{\omega\in\Omega}$, and let $\bar{M}\subseteq M$ be the set of messages in the support of $\{\mu(\cdot|s)\}_{s\in\tilde{S}}$. We will prove the result under Condition 1 and Condition 2 separately. For Condition 1 (resp., Condition 2), we first construct a signal structure $\bar{\pi}_1$ and a pure and separating messaging strategy $\bar{\mu}_1$ (resp., $\bar{\pi}_2$ and $\bar{\mu}_2$). We then show that $(\bar{\mu}_1;\alpha)$ (resp., $(\bar{\mu}_2;\alpha)$) is a PBE of the stage-2 communication game under $\bar{\pi}_1$ (resp., $\bar{\pi}_2$). Since the Receiver's strategy is unchanged at α, her best response condition (i.e., (R-IC)) is already satisfied, so we have to only verify the Sender's best response in (S-IC). After verifying the PBE, we show that $W(\bar{\pi}_1,\bar{\mu}_1;\alpha)\geq W(\pi,\mu;\alpha)$ (resp., $W(\bar{\pi}_2,\bar{\mu}_2;\alpha)\geq W(\pi,\mu;\alpha)$). Note that under both $(\bar{\pi}_1,\bar{\mu}_1)$ and $(\bar{\pi}_2,\bar{\mu}_2)$ below, the sets of supported messages in equilibrium will remain to be $\bar{M}$ (i.e., same as under (π,μ)). For any off-the-equilibrium message $m\notin\bar{M}$ under $(\bar{\pi}_1,\bar{\mu}_1)$ and $(\bar{\pi}_2,\bar{\mu}_2)$, the Receiver will hold the same belief as upon receiving the same message under (π,μ). Throughout, let $\Gamma:\bar{M}\rightrightarrows\tilde{S}$ be a correspondence, where $s\in\Gamma(m)$ if and only if m is in the support of $\mu(\cdot|s)$.

When Condition 1 holds:

Let $\underline{s}(m):=\arg\min_{s\in\Gamma(m)}c(m|s)$,[23] and let $\gamma_1:\bar{M}\to S$ be a function that satisfies the following two properties:

1. For each $m\in\bar{M}$, $c(m'|\gamma_1(m))=c(m'|\underline{s}(m))\ \forall m'\in M$.

2. $\gamma_1(m)=\gamma_1(m')$ if and only if $m=m'$.

Assumption 1 assures that γ_1 exists: if $\underline{s}(m)=\underline{s}(m')$ for some $m\neq m'$, then $\gamma_1(m)$ and $\gamma_1(m')$ are simply different signals with the same base signal. Let $\bar{S}_1:=\left\{\gamma_1(m)\,|m\in\bar{M}\right\}$, which is the support of $\bar{\pi}_1$ (to be defined). Since γ_1 is injective but not necessarily surjective, we redefine the codomain of γ_1 to be its range $\bar{S}_1$, so γ_1 is now bijective and thus, its inverse $\gamma_1^{-1}:\bar{S}_1\to\bar{M}$ exists (i.e., for any $m\in\bar{M}$, $\gamma_1^{-1}(\gamma_1(m))=m$). The signal structure $\bar{\pi}_1$ is defined as follows: for any Borel set $\hat{S}\subseteq S$,

$$\bar{\pi}_1\left(\hat{S}|\omega\right):=\int_{s\in S}\mu\left(\left\{\gamma_1^{-1}(s')\,|s'\in\hat{S}\cap\bar{S}_1\right\}|s\right)d\pi(s|\omega)\qquad\forall\omega\in\Omega.$$

Notice that $\left\{\gamma_1^{-1}(s')\,|s'\in S\cap\bar{S}_1\right\}=\bar{M}$. This implies that $\bar{\pi}_1(S|\omega)=1\ \forall\omega\in\Omega$, so $\bar{\pi}_1$ is a valid signal structure. Next, for any $s\in\bar{S}_1$, define $\bar{\mu}_1(\cdot|s)$ to be Dirac at $m=\gamma_1^{-1}(s)$. The

[23] If there are multiple minimums, just choose one of them.

messaging strategy after $s \notin \bar{S}_1$ is irrelevant because it is not supported by $\bar{\pi}_1(\cdot|\omega)$ for any $\omega \in \Omega$. $\bar{\mu}_1$ is clearly a valid messaging strategy that is fully separating and pure.

We verify that $(\bar{\mu}_1; \alpha)$ is a PBE under $\bar{\pi}_1$. Recall that the support of messages under $(\bar{\pi}_1, \bar{\mu}_1)$ is still $\bar{M}$. Under $(\bar{\pi}_1, \bar{\mu}_1)$, conditional on a state $\omega \in \Omega$, the probability measure of any arbitrary Borel set $\hat{M} \subseteq \bar{M}$ is

$$\begin{aligned} P_{\bar{\pi}_1,\bar{\mu}_1}\left(\hat{M}|\omega\right) =& \int_{s\in S} \bar{\mu}_1\left(\hat{M}|s\right) d\bar{\pi}_1(s|\omega) \qquad (16) \\ =& \int_{s\in\left\{\gamma_1(m)|m\in\hat{M}\right\}} d\bar{\pi}_1(s|\omega) \\ =& \int_{s'\in S}\int_{m\in\hat{M}} d\mu(m|s')\, d\pi(s'|\omega) \\ =& P_{\pi,\mu}\left(\hat{M}|\omega\right), \qquad (17) \end{aligned}$$

where $P_{\pi,\mu}\left(\hat{M}|\omega\right)$ is the probability measure of $\hat{M}$ conditional on ω under (π, μ). Therefore, for any $m \in \bar{M}$, $\rho_{\bar{\pi}_1,\bar{\mu}_1}(\cdot|m) = \rho_{\pi,\mu}(\cdot|m)$, which implies that the distribution of the Receiver's actions after any message $m \in \bar{M}$ is the same between $(\bar{\pi}_1, \bar{\mu}_1)$ and (π, μ).

Under Condition 1, we can write $v(a, \omega) = \nu(a)\ \forall\omega$. We abuse notation as previously and let ν denote the expected payoff as well.[24] For every $m \in \bar{M}$, the Sender's incentive compatibility condition upon seeing signal $\underline{s}(m)$ under $(\pi, \mu; \alpha)$ (see (S-IC)) implies that for any $m' \in M$,

$$\begin{aligned} & \nu(\alpha(\cdot|\rho_{\pi,\mu}(\cdot|m))) - \nu(\alpha(\cdot|\rho_{\pi,\mu}(\cdot|m'))) \geq c(m|\underline{s}(m)) - c(m'|\underline{s}(m)) \\ \Longrightarrow\ & \nu(\alpha(\cdot|\rho_{\bar{\pi}_1,\bar{\mu}_1}(\cdot|m))) - \nu(\alpha(\cdot|\rho_{\bar{\pi}_1,\bar{\mu}_1}(\cdot|m'))) \geq c(m|\gamma_1(m)) - c(m'|\gamma_1(m)) \qquad (18) \end{aligned}$$

This implies that $\forall s \in \bar{S}_1$,

$$\nu\left(\alpha\left(\cdot|\rho_{\bar{\pi}_1,\bar{\mu}_1}\left(\cdot|\gamma_1^{-1}(s)\right)\right)\right) - \nu(\alpha(\cdot|\rho_{\bar{\pi}_1,\bar{\mu}_1}(\cdot|m'))) \geq c\left(\gamma_1^{-1}(s)|s\right) - c(m'|s) \quad \forall m' \in M. \qquad (19)$$

Since $\bar{\mu}_1(\cdot|s)$ is Dirac on $\gamma_1^{-1}(s)\ \forall s \in \bar{S}_1$, $\bar{\mu}_1(\cdot|s)$ satisfies the Sender's best response constraint in (S-IC) $\forall s \in \bar{S}_1$. Therefore, $(\bar{\mu}_1; \alpha)$ is a PBE under $\bar{\pi}_1$.

[24] i.e., for any $\alpha(\cdot|\rho) \in \Delta A$, $\nu(\alpha(\cdot|\rho)) = \int_{a\in A} \nu(a)\, d\alpha(a|\rho)$.

To check that $W\left(\bar{\pi}_1, \bar{\mu}_1; \alpha\right) \geq W\left(\mu, \mu; \alpha\right)$, note first that

$$\begin{aligned} V\left(\bar{\pi}_1, \bar{\mu}_1; \alpha\right) &= \sum_{\omega\in\Omega} \beta^o\left(\omega\right) \int_{m\in\bar{M}} \nu\left(\alpha\left(\cdot|\rho_{\bar{\pi}_1,\bar{\mu}_1}\left(\cdot|m\right)\right)\right) dP_{\bar{\pi}_1,\bar{\mu}_1}\left(m|\omega\right) \\ &= \sum_{\omega\in\Omega} \beta^o\left(\omega\right) \int_{m\in\bar{M}} \nu\left(\alpha\left(\cdot|\rho_{\pi,\mu}\left(\cdot|m\right)\right)\right) dP_{\pi,\mu}\left(m|\omega\right) \\ &=V\left(\pi, \mu; \alpha\right). \end{aligned}$$

Next

$$\begin{aligned} C\left(\pi, \mu\right) &= \sum_{\omega\in\Omega} \beta^o\left(\omega\right) \int_{s\in S} \int_{m\in\bar{M}} c\left(m|s\right) d\mu\left(m|s\right) d\pi\left(s|\omega\right) \\ &\geq \sum_{\omega\in\Omega} \beta^o\left(\omega\right) \int_{s\in S} \int_{m\in\bar{M}} c\left(m|\gamma_1\left(m\right)\right) d\mu\left(m|s\right) d\pi\left(s|\omega\right) && (20) \\ &= \sum_{\omega\in\Omega} \beta^o\left(\omega\right) \int_{m\in\bar{M}} c\left(m|\gamma_1\left(m\right)\right) dP_{\pi,\mu}\left(m|\omega\right) && (21) \\ &= \sum_{\omega\in\Omega} \beta^o\left(\omega\right) \int_{m\in\bar{M}} c\left(m|\gamma_1\left(m\right)\right) dP_{\bar{\pi}_1,\bar{\mu}_1}\left(m|\omega\right) && (22) \\ &=C\left(\bar{\pi}_1, \bar{\mu}_1\right) \end{aligned}$$

The inequality in line (20) follows from noting that m is in the support of $\mu\left(\cdot|s\right)$ only if $s \in \Gamma\left(m\right)$, and $c\left(m|s\right) \leq c\left(m|\underline{s}\left(m\right)\right) = c\left(m|\gamma_1\left(m\right)\right)\ \forall s \in \Gamma\left(m\right)$. The equality from line (21) to line (22) follows from lines (16) to (17). Therefore, $W\left(\bar{\pi}_1, \bar{\mu}_1; \alpha\right) \geq W\left(\mu, \mu; \alpha\right)$.

When Condition 2 holds:

Let $\Lambda\left(\cdot|m\right) \in \Delta S$ be the regular conditional probability measure over the signal space when conditioned on the Receiver receiving message m under $\left(\pi, \mu\right)$.[25] By the martingale property of Bayesian posteriors,

$$\rho_{\pi,\mu}\left(\cdot|m\right) = \int_{s\in\Gamma(m)} \sigma_\pi\left(\cdot|s\right) d\Lambda\left(s|m\right) \qquad \forall m \in \bar{M}. \tag{23}$$

Define the function $\gamma_2 : \bar{M} \to S$, where for each $m \in \bar{M}$:

[25] i.e., for any Borel sets $\hat{S} \subseteq S$ and $\hat{M} \subseteq M$,

$$\sum_{\omega\in\Omega} \beta^o\left(\omega\right) \int_{s\in\hat{S}} \int_{m\in\hat{M}} d\mu\left(m|s\right) d\pi\left(s|\omega\right) = \sum_{\omega'\in\Omega} \beta^o\left(\omega'\right) \int_{m\in\hat{M}} \Lambda\left(\hat{S}|m\right) dP_{\pi,\mu}\left(m|\omega'\right)$$

1. $c\left(m|\gamma_2\left(m\right)\right) \leq \int_{s\in\Gamma(m)} c\left(m|s\right) d\Lambda\left(s|m\right)$.
2. $c\left(m|\gamma_2\left(m\right)\right) - \int_{s\in S} c\left(m|s\right) d\Lambda\left(s|m\right) \leq c\left(m'|\gamma_2\left(m\right)\right) - \int_{s\in S} c\left(m'|s\right) d\Lambda\left(s|m\right) \;\; \forall m' \in M$ such that $\int_{s\in S} c\left(m'|s\right) d\Lambda\left(s|m\right) < \infty$.
3. $c\left(m'|\gamma_2\left(m\right)\right) = \infty \;\; \forall m' \in M$ such that $\int_{s\in S} c\left(m'|s\right) d\Lambda\left(s|m\right) = \infty$.
4. $\gamma_2\left(m'\right) = \gamma_2\left(m''\right)$ if and only if $m' = m''$.

Since $\left(\pi, \mu; \alpha\right)$ is an equilibrium, $\int_{s\in S} c\left(m|s\right) d\Lambda\left(s|m\right) < \infty \; \forall m \in \bar{M}$, so a γ_2 function satisfying the first three properties exists when Condition 2 is satisfied. By Assumption 1, there are infinitely many signals with the same messaging costs, so the last property for γ_2 is also readily satisfied. Let $\bar{S}_2 := \left\{\gamma_2\left(m\right)|m \in \bar{M}\right\}$, which is the support of $\bar{\pi}_2$ (to be defined). Analogous to γ_1 above, we redefine the codomain of γ_2 to be its range $\bar{S}_2$ so that γ_2 is bijective and, hence, its inverse function $\gamma_2^{-1} : \bar{S}_2 \rightarrow \bar{M}$ exists (i.e., $\gamma_2^{-1}\left(\gamma_2\left(m\right)\right) = m \; \forall m \in \bar{M}$). The signal structure $\bar{\pi}_2$ is defined as follows: for any Borel set $\hat{S} \subseteq S$,

$$\bar{\pi}_2\left(\hat{S}|\omega\right) := \int_{s\in S} \mu\left(\left\{\gamma_2^{-1}\left(s'\right)|s' \in \hat{S} \cap \bar{S}_2\right\}|s\right) d\pi\left(s|\omega\right) \quad \forall \omega \in \Omega. \tag{24}$$

Analogous to $\bar{\pi}_1$ above, it is readily seen that $\bar{\pi}_2\left(S|\omega\right) = 1 \; \forall \omega \in \Omega$; therefore, $\bar{\pi}_2$ is a valid signal structure. Let $\bar{\mu}_2\left(\cdot|s\right)$ be Dirac on $m = \gamma_2^{-1}\left(s\right) \; \forall s \in \bar{S}_2$; therefore, $\bar{\mu}_2$ is a fully separating and pure messaging strategy.

We verify that $\left(\bar{\mu}_2; \alpha\right)$ is a PBE under $\bar{\pi}_2$ next. First, note that $\sigma_{\bar{\pi}_2}\left(\cdot|s\right) = \rho_{\bar{\pi}_2,\bar{\mu}_2}\left(\cdot|\gamma_2^{-1}\left(s\right)\right)$ $\forall s \in \bar{S}_2$. Moreover, by an argument that is analogous to lines (16) to (17), the probability measures of any arbitrary Borel set $\hat{M} \subseteq \bar{M}$ when conditional on a $\omega \in \Omega$ are the same under $\left(\pi, \mu\right)$ and $\left(\bar{\pi}_2, \bar{\mu}_2\right)$ – i.e.,

$$P_{\bar{\pi}_2,\bar{\mu}_2}\left(\hat{M}|\omega\right) = P_{\pi,\mu}\left(\hat{M}|\omega\right). \tag{25}$$

This implies that $\rho_{\bar{\pi}_2,\bar{\mu}_2}\left(\cdot|m\right) = \rho_{\pi,\mu}\left(\cdot|m\right) \; \forall m \in \bar{M}$. Therefore, for any $\alpha \in \Delta A$ and $s \in \bar{S}_2$,

$$\begin{aligned} v\left(\alpha, \sigma_{\bar{\pi}_2}\left(\cdot|s\right)\right) &= \sum_{\omega\in\Omega} \rho_{\bar{\pi}_2,\bar{\mu}_2}\left(\omega|\gamma_2^{-1}\left(s\right)\right) v\left(\alpha, \omega\right) \\ &= \sum_{\omega\in\Omega} \rho_{\pi,\mu}\left(\omega|\gamma_2^{-1}\left(s\right)\right) v\left(\alpha, \omega\right) \qquad (26) \\ &= \sum_{\omega\in\Omega} \int_{s'\in\Gamma\left(\gamma_2^{-1}(s)\right)} \sigma_\pi\left(\omega|s'\right) d\Lambda\left(s'|\gamma_2^{-1}\left(s\right)\right) v\left(\alpha, \omega\right) \qquad (27) \\ &= \int_{s'\in\Gamma\left(\gamma_2^{-1}(s)\right)} v\left(\alpha, \sigma_\pi\left(\cdot|s'\right)\right) d\Lambda\left(s'|\gamma_2^{-1}\left(s\right)\right), \qquad (28) \end{aligned}$$

where the equality in (27) follows from (23). Next, we ease notation and denote, for each $m \in \bar{M}$, $\alpha_m(\cdot) = \alpha(\cdot|\rho_{\bar{\pi}_2,\bar{\mu}_2}(\cdot|m)) = \alpha(\cdot|\rho_{\pi,\mu}(\cdot|m))$. Since $(\pi, \mu; \alpha)$ is an equilibrium, the Sender's best response constraint (S-IC) implies that

$$\forall m \in \bar{M} \text{ and } s \in \Gamma(m), \quad v\Big(\alpha_m, \sigma_\pi(\cdot|s)\Big) - v\Big(\alpha_{m'}, \sigma_\pi(\cdot|s)\Big) \geq c(m|s) - c(m'|s) \quad \forall m' \neq m. \tag{29}$$

Fix any $s \in \bar{S}_2$. Under $\bar{\mu}_2$, the Sender reports $m = \gamma_2^{-1}(s) \in \bar{M}$ with probability one. To establish that the Sender's best response constraint (S-IC) is satisfied under $(\bar{\pi}_2, \bar{\mu}_2; \alpha)$, we have to show that $\forall m' \neq \gamma_2^{-1}(s)$,

$$v\left(\alpha_{\gamma_2^{-1}(s)}, \sigma_{\bar{\pi}_2}(\cdot|s)\right) - v\left(\alpha_{m'}, \sigma_{\bar{\pi}_2}(\cdot|s)\right) \geq c\left(\gamma_2^{-1}(s)|s\right) - c(m'|s) \tag{30}$$

First, suppose that m' is such that $\int_{s' \in \Gamma\left(\gamma_2^{-1}(s)\right)} c(m'|s')\, d\Lambda\left(s'|\gamma_2^{-1}(s)\right) = \infty$. Property 3 of γ_2 implies that $c(m'|s) = \infty$. Given that v is bounded and $c\left(\gamma_2^{-1}(s)|s\right) < \infty$ (from Property 1 of γ_2), (30) is satisfied. Next, suppose that m' is such that $\int_{s' \in \Gamma\left(\gamma_2^{-1}(s)\right)} c(m'|s')\, d\Lambda\left(s'|\gamma_2^{-1}(s)\right) < \infty$. We have

$$\begin{aligned}
& v\left(\alpha_{\gamma_2^{-1}(s)}, \sigma_{\bar{\pi}_2}(\cdot|s)\right) - v\left(\alpha_{m'}, \sigma_{\bar{\pi}_2}(\cdot|s)\right) \\
=& \int_{s' \in \Gamma\left(\gamma_2^{-1}(s)\right)} \left[v\left(\alpha_{\gamma_2^{-1}(s)}, \sigma_\pi(\cdot|s')\right) - v\left(\alpha_{m'}, \sigma_\pi(\cdot|s')\right)\right] d\Lambda\left(s'|\gamma_2^{-1}(s)\right) && (31) \\
\geq& \int_{s' \in \Gamma\left(\gamma_2^{-1}(s)\right)} \left[c\left(\gamma_2^{-1}(s)|s'\right) - c(m'|s')\right] d\Lambda\left(s'|\gamma_2^{-1}(s)\right) && (32) \\
\geq& c\left(\gamma_2^{-1}(s)|s\right) - c(m'|s), && (33)
\end{aligned}$$

where the equality in (31) follows from (28), the inequality in (32) follows from (29), and the inequality in (33) follows from Property 2 of γ_2. Therefore, we have checked the Sender's best response constraint (S-IC) is satisfied under $(\bar{\pi}_2, \bar{\mu}_2; \alpha)$, so $(\bar{\mu}_2; \alpha)$ is a PBE under $\bar{\pi}_2$.

Next, by the Radon-Nikodym theorem, σ_π in (1) defines the Sender's posterior almost everywhere, so for any integrable function $g : S \times \Omega \to \mathbb{R}$,

$$\sum_{\omega' \in \Omega} \beta^o(\omega') \int_{s \in S} \sigma_\pi(\omega|s)\, g(s, \omega)\, d\pi(s|\omega') = \beta^o(\omega) \int_{s \in S} g(s, \omega)\, d\pi(s|\omega) \tag{34}$$

Therefore,

$$\begin{aligned}
V(\pi,\mu;\alpha) &= \sum_{\omega'\in\Omega} \beta^o(\omega') \int_{s\in S} \int_{m\in\bar{M}} v(\alpha_m, \sigma_\pi(\cdot|s))\, d\mu(m|s)\, d\pi(s|\omega') \\
&= \sum_{\omega'\in\Omega} \beta^o(\omega') \int_{s\in S} \int_{m\in\bar{M}} \sum_{\omega\in\Omega} \sigma_\pi(\omega|s)\, v(\alpha_m,\omega)\, d\mu(m|s)\, d\pi(s|\omega') \\
&= \sum_{\omega\in\Omega}\sum_{\omega'\in\Omega} \beta^o(\omega') \int_{s\in S} \sigma_\pi(\omega|s) \int_{m\in\bar{M}} v(\alpha_m,\omega)\, d\mu(m|s)\, d\pi(s|\omega') \\
&= \sum_{\omega\in\Omega} \beta^o(\omega) \int_{s\in S} \int_{m\in\bar{M}} v(\alpha_m,\omega)\, d\mu(m|s)\, d\pi(s|\omega) \qquad (35) \\
&= \sum_{\omega\in\Omega} \beta^o(\omega) \int_{m\in\bar{M}} v(\alpha_m,\omega)\, dP_{\pi,\mu}(m|\omega), \qquad (36)
\end{aligned}$$

where the equality in (35) follows from (34). By the same argument,

$$V(\bar{\pi}_2,\bar{\mu}_2;\alpha) = \sum_{\omega\in\Omega} \beta^o(\omega) \int_{m\in\bar{M}} v(\alpha_m,\omega)\, dP_{\bar{\pi}_2,\bar{\mu}_2}(m|\omega). \qquad (37)$$

By (25), $P_{\bar{\pi}_2,\bar{\mu}_2}(\cdot|\omega) = P_{\pi,\mu}(\cdot|\omega)\ \forall\omega$, so (36) and (37) jointly imply that $V(\bar{\pi}_2,\bar{\mu}_2;\alpha) = V(\pi,\mu;\alpha)$. Next,

$$\begin{aligned}
C(\bar{\pi}_2,\bar{\mu}_2) &= \sum_{\omega'\in\Omega} \beta^o(\omega') \int_{s\in S} \int_{m\in\bar{M}} c(m|s)\, d\bar{\mu}_2(m|s)\, d\bar{\pi}_2(s|\omega') \qquad (38) \\
&= \sum_{\omega'\in\Omega} \beta^o(\omega') \int_{m\in\bar{M}} [c(m|\gamma_2(m))]\, dP_{\bar{\pi}_2,\bar{\mu}_2}(m|\omega) \\
&\leq \sum_{\omega'\in\Omega} \beta^o(\omega') \int_{m\in\bar{M}} \left[\int_{s\in\Gamma(m)} c(m|s)\, d\Lambda(s|m)\right] dP_{\bar{\pi}_2,\bar{\mu}_2}(m|\omega) \qquad (39) \\
&= \sum_{\omega'\in\Omega} \beta^o(\omega') \int_{m\in\bar{M}} \left[\int_{s\in\Gamma(m)} c(m|s)\, d\Lambda(s|m)\right] dP_{\pi,\mu}(m|\omega) \qquad (40) \\
&= \sum_{\omega'\in\Omega} \beta^o(\omega') \int_{s'\in S} \int_{m\in\bar{M}} \left[\int_{s\in\Gamma(m)} c(m|s)\, d\Lambda(s|m)\right] d\mu(m|s')\, d\pi(s'|\omega') \\
&= \sum_{\omega\in\Omega} \beta^o(\omega) \int_{s'\in S} \int_{m\in\bar{M}} c(m|s')\, d\mu(m|s')\, d\pi(s'|\omega) \qquad (41) \\
&= C(\pi,\mu). \qquad (42)
\end{aligned}$$

The inequality in (39) follows from the first property of γ_2, the equality in (40) follows from (25), and the equality in (41) follows from the definition of $\Gamma(m)$ that it is the set of signals s in which m is in the support of $\mu(\cdot|s)$. Since $V(\bar{\pi}_2,\bar{\mu}_2;\alpha) = V(\pi,\mu;\alpha)$ and

$C(\bar{\pi}_2, \bar{\mu}_2; \alpha) \leq C(\pi, \mu; \alpha)$, we have $W(\bar{\pi}_2, \bar{\mu}_2; \alpha) \geq W(\pi, \mu; \alpha)$. □

Proof of Proposition 2.

Proof. Consider a Sender-optimal equilibrium $(\pi, \mu; \alpha)$, where the set of signals supported by $\{\pi(\cdot|\omega)\}_{\omega\in\Omega}$ is $\bar{S}$, and $\mu(\cdot|s)$ is Dirac on $m = \gamma^{-1}(s)$, where $\gamma : \bar{M} \to \bar{S}$ is a bijection, and $\bar{M} = \left\{\gamma^{-1}(s) | s \in \bar{S}\right\} \subseteq M$ is the set of messages in the support of $\{\mu(\cdot|s)\}_{s\in\bar{S}}$. From Proposition 1, $(\pi, \mu; \alpha)$ exists. If $\left|\bar{M}\right| \leq |\Omega|$, we are done, so we suppose that $\left|\bar{M}\right| > |\Omega|$. Let $Q_{\pi,\mu}(\cdot) \in \Delta\bar{M}$ be the equilibrium probability measure of the messages under (π, μ).[26]We introduce a few notations. Note first that since $\mu(\cdot|\gamma(m))$ is Dirac on m, $\sigma_\pi(\cdot|\gamma(m)) = \rho_{\pi,\mu}(\cdot|m)\ \forall m \in \bar{M}$. Let $\mathcal{B} = \left\{\sigma_\pi(\cdot|\gamma(m)) | m \in \bar{M}\right\} \subset \Delta\Omega$, and $\forall m \in \bar{M}$,

$$\begin{aligned} y_m &:= v\left(\alpha\left(\cdot|\rho_{\pi,\mu}(\cdot|m)\right), \sigma_\pi(\cdot|\gamma(m))\right) - c\left(m|\gamma(m)\right) \ \in\ \mathbb{R}, \\ z_m &:= \left\{\sigma_\pi(\cdot|\gamma(m)), y_m\right\} \ \in\ \mathcal{B} \times \mathbb{R} \ \subset\ \mathbb{R}^{|\Omega|}. \end{aligned}$$

Let $\mathcal{Z} := \left\{z_m | m \in \bar{M}\right\} \subset \mathbb{R}^{|\Omega|}$. For any set $\mathcal{A}$, let $co(\mathcal{A})$ denote its convex hull.

Notice that $W(\pi, \mu; \alpha) = \int_{m\in\bar{M}} y_m dQ_{\pi,\mu}(m)$ and $\beta^o = \int_{m\in\bar{M}} \sigma_\pi(\cdot|\gamma(m))\, dQ_{\pi,\mu}(m)$; therefore, $\{\beta^o, W(\pi, \mu; \alpha)\} \in co(\mathcal{Z})$.[27] Moreover, $co(\mathcal{B}) \subset \Delta\Omega$. Define the function $Y : co(\mathcal{B}) \to \mathbb{R}$, where $Y(\sigma_\pi(\cdot|\gamma(m))) = y_m\ \forall m \in \bar{M}$ and $Y(\beta) = \xi$ for any $\beta \notin \mathcal{B}$, where ξ is any fixed finite value that is strictly less than $\min_{m\in\bar{M}} y_m$.[28] Let $hyp(Y)$ be the hypograph of Y – i.e., $hyp(Y) = \{\{\beta, y\} | \beta \in co(\mathcal{B}), y \leq Y(\beta)\} \subset \mathbb{R}^{|\Omega|}$. Recall that v is bounded above by $\bar{v}$. Since $Y(\beta) \leq \bar{v}\ \forall \beta \in co(\mathcal{B})$, $hyp(Y)$ is path-connected and, thus, connected. By the Fenchel-Bunt theorem (see Hiriart-Urruty and Lemaréchal (2012), pp. 30, Theorem 1.3.7), any element in $co(hyp(Y))$ can be written as a convex combination of at most $|\Omega|$ elements

[26] i.e., for any Borel set $\hat{M} \subseteq \bar{M}$,

$$Q_{\pi,\mu}\left(\hat{M}\right) = \sum_{\omega\in\Omega} \beta^o(\omega) \int_{s\in\bar{S}} \mu\left(\hat{M}|s\right) d\pi(s|\omega).$$

[27] At this point, the well-known Carathéodory theorem (see Hiriart-Urruty and Lemaréchal (2012), pp. 29, Theorem 1.3.6) will imply that $\{\beta^o, W(\pi, \mu; \alpha)\}$ can be written as a convex combination of $|\Omega| + 1$ or less elements from $\mathcal{Z}$. However, we want to prove that it can be written as a convex combination of only $|\Omega|$ or less elements from $\mathcal{Z}$.

[28] Note that y_m is the Sender's utility after message $m \in \bar{M}$. Since m is supported in the equilibrium $(\pi, \mu; \alpha)$, y_m must be greater than $-\infty$.

in $hyp\left(Y\right)$. Since $co\left(\mathcal{Z}\right)\subset co\left(hyp\left(Y\right)\right)$, there exists a set $\tilde{\mathcal{Z}}\subset co\left(hyp\left(Y\right)\right)$, with $\left|\tilde{\mathcal{Z}}\right|\leq\left|\Omega\right|$, and $\tilde{\tau}\in\Delta\tilde{\mathcal{Z}}$ such that $\sum_{\tilde{z}\in\tilde{\mathcal{Z}}}\tilde{\tau}\left(\tilde{z}\right)\tilde{z}=\left\{\beta^{o},W\left(\pi,\mu;\alpha\right)\right\}$.

We are left to show that $\tilde{\mathcal{Z}}\subset\mathcal{Z}$. Let $\tilde{\mathcal{Z}}=\left\{\tilde{z}_{1},\ldots,\tilde{z}_{K}\right\}$, where $K\leq\left|\Omega\right|$, and $\tilde{z}_{k}=\left\{\tilde{\beta}_{k},\tilde{y}_{k}\right\}$, with $\tilde{\beta}_{k}\in co\left(\mathcal{B}\right)\subseteq\Delta\Omega$, $\tilde{y}_{k}\in\mathbb{R}\ \forall k=1,\ldots,K$. Since $\tilde{\beta}_{k}\in co\left(\mathcal{B}\right)\ \forall k$, by the Carathéodory theorem, there exists a finite set $\hat{\mathcal{B}}=\left\{\hat{\beta}_{1},\ldots,\hat{\beta}_{J}\right\}\subset\mathcal{B}$ such that for every $k=1,2,\ldots,K$, there exists $\delta_{k}\in\Delta\hat{\mathcal{B}}$ such that $\sum_{j=1}^{J}\delta_{k}\left(\hat{\beta}_{j}\right)\hat{\beta}_{j}=\tilde{\beta}_{k}$. We take the convention that if $\tilde{\beta}_{k}\in\mathcal{B}$, then we let $\tilde{\beta}_{k}\in\hat{\mathcal{B}}$ with $\delta_{k}\left(\tilde{\beta}_{k}\right)=1$.[29] Now, suppose, for a contradiction, that $\tilde{\mathcal{Z}}\not\subset\mathcal{Z}$. When $\tilde{z}_{k}=\left\{\tilde{\beta}_{k},\tilde{y}_{k}\right\}\notin\mathcal{Z}$, it must imply that either (i) $\tilde{\beta}_{k}\in\mathcal{B}$ but $\tilde{y}_{k}<Y\left(\tilde{\beta}_{k}\right)$, or (ii) $\tilde{\beta}_{k}\notin\mathcal{B}$, which implies that $\tilde{y}_{k}=\xi<\sum_{j=1}^{J}\delta_{k}\left(\hat{\beta}_{j}\right)Y\left(\hat{\beta}_{j}\right)$. Therefore, both cases will imply that

$$\begin{aligned}W\left(\pi,\mu;\alpha\right)&=\sum_{k=1}^{K}\tilde{\tau}\left(\tilde{z}_{k}\right)\tilde{y}_{k}\\&<\sum_{k=1}^{K}\tilde{\tau}\left(\tilde{z}_{k}\right)\sum_{j=1}^{J}\delta_{k}\left(\hat{\beta}_{j}\right)Y\left(\hat{\beta}_{j}\right)=\Psi.\end{aligned}$$

We will show that Ψ is attainable for the Sender in equilibrium, which then contradicts $\left(\pi,\mu;\alpha\right)$ being a Sender-optimal equilibrium. Let $\hat{\tau}\in\Delta\Delta\Omega$ be a probability measure with support on $\hat{\mathcal{B}}$ and $\hat{\tau}\left(\hat{\beta}_{j}\right)=\sum_{k=1}^{K}\tilde{\tau}\left(\tilde{z}_{k}\right)\delta_{k}\left(\hat{\beta}_{j}\right)$. Notice that

$$\begin{aligned}\sum_{j=1}^{J}\hat{\tau}\left(\hat{\beta}_{j}\right)\hat{\beta}_{j}=&\sum_{k=1}^{K}\tilde{\tau}\left(\tilde{z}_{k}\right)\sum_{j=1}^{J}\delta_{k}\left(\hat{\beta}_{j}\right)\hat{\beta}_{j}\\=&\sum_{k=1}^{K}\tilde{\tau}\left(\tilde{z}_{k}\right)\tilde{\beta}_{k}\quad=\quad\beta^{o}.\end{aligned}$$

This implies that there exists a set of signals $\hat{S}=\left\{s_{1},s_{2},\ldots,s_{j}\right\}\subset S$ and a measurable map $\hat{\pi}:\Omega\rightarrow\Delta\hat{S}$ such that for each j and $\omega\in\Omega$, $\hat{\beta}_{j}\left(\omega\right)=\frac{\hat{\pi}\left(s_{j}|\omega\right)\beta^{o}\left(\omega\right)}{\sum_{\omega'\in\Omega}\hat{\pi}\left(s_{j}|\omega'\right)\beta^{o}\left(\omega'\right)}$, $\sum_{j=1}^{J}\hat{\pi}\left(s_{j}|\omega\right)=1$, and $\hat{\tau}\left(\hat{\beta}_{j}\right)=\sum_{\omega\in\Omega}\beta^{o}\left(\omega\right)\hat{\pi}\left(s_{j}|\omega\right)$ (see Proposition 1 in KG). Since $\hat{\mathcal{B}}\subset\mathcal{B}$, for each $j\in\left\{1,2,\ldots,J\right\}$, there exists a function $g:\left\{1,2,\ldots,J\right\}\rightarrow\bar{M}$ such that $\hat{\beta}_{j}\left(\cdot\right)=\sigma_{\pi}\left(\cdot|\gamma\left(g\left(j\right)\right)\right)$. Consider the signal structure with the same probability measure as $\hat{\pi}$ with $s_{j}=\gamma\left(g\left(j\right)\right)$; we abuse notation and let $\hat{\pi}$ denote this signal structure.[30] By construction, $\sigma_{\hat{\pi}}\left(\cdot|s_{j}\right)=$

[29] By the Carathéodory theorem, for each k, there exists $\hat{\mathcal{B}}^{k}=\left\{\hat{\beta}_{1}^{k},\ldots,\hat{\beta}_{|\Omega|+1}^{k}\right\}\subset\mathcal{B}$ and $\hat{\delta}_{k}\in\Delta\hat{\mathcal{B}}^{k}$ such that $\sum_{j=1}^{|\Omega|+1}\hat{\delta}_{k}\left(\hat{\beta}_{j}^{k}\right)\hat{\beta}_{j}^{k}=\tilde{\beta}_{k}$; we take the convention that if $\tilde{\beta}_{k}\in\mathcal{B}$, then $\hat{\mathcal{B}}^{k}=\left\{\tilde{\beta}_{k}\right\}$ and $\hat{\delta}_{k}\left(\tilde{\beta}_{k}\right)=1$. Therefore, $\hat{\mathcal{B}}=\cup_{k}\hat{\mathcal{B}}^{k}$; and $\delta_{k}\left(\beta\right)=\hat{\delta}_{k}\left(\beta\right)$ if $\beta\in\hat{\mathcal{B}}^{k}$, and zero otherwise.

[30] The abuse of notation arises as a signal structure is defined as a measurable map from Ω to ΔS, whereas

$\sigma_\pi(\cdot|\gamma(g(j)))$. Consider the messaging strategy $\hat{\mu}$, where for each $s_j \in \hat{S}$, $\hat{\mu}(\cdot|s_j)$ is Dirac on $g(j)$. Therefore, for each $m \in \hat{M} = \{g(1), g(2), \ldots, g(J)\}$, which is the set of messages supported by $\hat{\pi}$ and $\hat{\mu}$, $\rho_{\hat{\pi},\hat{\mu}}(\cdot|m) = \sigma_{\hat{\pi}}\left(\cdot|s_{g^{-1}(m)}\right) = \sigma_\pi(\cdot|\gamma(m))$. Thus, when the Sender's best response condition in (S-IC) is satisfied for $(\mu;\alpha)$ under π, it will also be satisfied for $(\hat{\mu};\alpha)$ under $\hat{\pi}$, so $(\hat{\mu};\alpha)$ is a PBE under $\hat{\pi}$. Moreover, $W(\hat{\pi},\hat{\mu};\alpha) = \sum_{j=1}^{J} \hat{\tau}\left(\hat{\beta}_j\right) Y\left(\hat{\beta}_j\right) = \Psi > W(\pi,\mu;\alpha)$, which then contradicts $(\pi,\mu;\alpha)$ being a Sender-optimal equilibrium. □

Proof of Theorem 1.

Proof. Let $\Phi(\tau,\alpha) = V^*(\tau,\alpha) - C^*(\tau,\alpha)$ and let $\sup\limits_{\tau\in\mathcal{D},\alpha\in\mathcal{A}^*} \Phi(\tau,\alpha) = W^*$. Since the Sender's payoff is always finite, $W^* < \infty$. Suppose that a Sender-optimal equilibrium exists. Let $(\pi,\mu;\alpha)$ be a Sender-optimal equilibrium, and let $\bar{S} \subset S$ and $\bar{M} \subset M$ be, respectively, the set of signals and messages supported in the equilibrium. By Proposition 1, it is without loss to assume that $\left|\bar{S}\right| = \left|\bar{M}\right| \leq |\Omega|$, and there exists a bijection $g : \bar{S} \to \bar{M}$ such that $\mu(\cdot|s)$ is Dirac on $g(s)$ $\forall s \in \bar{S}$. We denote $\bar{S} = \{s_1, s_2, \ldots, s_K\}$, where $K \leq |\Omega|$, and let $m_k = g(s_k)$. Let $\beta(\omega|s) = \frac{\pi(\{s\}|\omega)\beta^o(\omega)}{\sum_{\omega'\in\Omega}\pi(\{s\}|\omega')\beta^o(\omega')}$ $\forall \omega \in \Omega$, $s \in \bar{S}$. It is readily verified that $\sigma_\pi(\cdot|s_k) = \beta(\cdot|s_k)$ $\forall s_k \in \bar{S}$, and $\rho_{\pi,\mu}(\cdot|m_k) = \beta(\cdot|s_k)$ $\forall m_k \in \bar{M}$. Let $\delta_k = \sum_{\omega\in\Omega}\beta^o(\omega)\pi(s_k|\omega)$; the ex-ante probabilities of signal s_k and m_k are both δ_k. Therefore,

$$V(\pi,\mu;\alpha) = \sum_{k=1}^{K} \delta_k v(\alpha(\cdot|\beta(\cdot|s_k)), \beta(\cdot|s_k)) = V^*\left(\left\{\vec{\beta};\vec{\delta}\right\},\alpha\right),$$

where $\vec{\beta} = \{\beta(\cdot|s_1), \ldots, \beta(\cdot|s_K)\}$ and $\vec{\delta} = \{\delta_1, \ldots, \delta_K\}$, and

$$C(\pi,\mu) = \sum_{\omega\in\Omega} \beta^o(\omega) \int_{s\in S}\int_{m\in M} c(m|s)\, d\mu(m|s)\, d\pi(s|\omega) = \sum_{k=1}^{K} \delta_k c(m_k|s_k).$$

Moreover, since $(\mu;\alpha)$ is a PBE under π, $\forall k, k' = 1, 2, \ldots, K$,

$$v\Big(\alpha(\cdot|\beta(\cdot|s_k)), \beta(\cdot|s_k)\Big) - c(m_k|s_k) \geq v\Big(\alpha(\cdot|\beta(\cdot|s_{k'})), \beta(\cdot|s_k)\Big) - c(m_{k'}|s_k).$$

Therefore, $C(\pi,\mu) \geq C^*\left(\left\{\vec{\beta};\vec{\delta}\right\},\alpha\right)$, which implies that

$$W(\pi,\mu;\alpha) \leq \Phi\left(\left\{\vec{\beta};\vec{\delta}\right\},\alpha\right) \leq W^*. \tag{43}$$

$\hat{\pi}$ is a map from Ω to $\Delta\hat{S}$.

.

Next, fix any $\left\{\vec{\beta}';\vec{\delta}'\right\} \in \mathcal{D}$, with $\vec{\beta}' = \{\beta_1',\dots,\beta_J'\}$ and $\vec{\delta}' = \{\delta_1',\dots,\delta_J'\}$ and $J \leq |\Omega|$, and any $\alpha' \in \mathcal{A}^*$. Let $\bar{S}' = \{s_1',\dots,s_J'\}$ and $\bar{M}' = \{m_1',\dots,m_J'\}$ be a solution to program (6) for $\left(\left\{\vec{\beta}';\vec{\delta}'\right\},\alpha'\right)$ subject to constraint (7). Consider the signal structure π' which puts point mass over $\bar{S}'$: for each $\omega \in \Omega$ and $j \in \{1,\dots,J\}$, set $\pi'(s_j|\omega) = \frac{\beta_j'(\omega)\delta_j}{\beta^o(\omega)}$. Set $\mu'(\cdot|s_j)$ to be Dirac on m_j $\forall j = 1,\dots,J$. It is readily verified that $\rho_{\pi',\mu'}(\cdot|s_j) = \rho_{\pi',\mu'}(\cdot|m_j) = \beta_j'(\cdot)$ $\forall j = 1,\dots,J$. Since $\bar{S}'$ and $\bar{M}'$ satisfied constraint (7), the Sender's best response constraint in (S-IC) is satisfied. This implies that $(\mu';\alpha')$ is a PBE under π' , and it is readily verified that $W(\pi',\mu';\alpha') = \Phi\left(\left\{\vec{\beta}';\vec{\delta}'\right\},\alpha'\right)$. Therefore, there exists an equilibrium strategy profile $(\pi',\mu';\alpha')$ that achieves $\Phi\left(\left\{\vec{\beta}';\vec{\delta}'\right\},\alpha'\right)$. This implies that if program $\mathcal{P}$ has no solution, the Sender-optimal equilibrium cannot exist. Conversely, if program $\mathcal{P}$ has a solution $\tau^* = \left\{\vec{\beta}^*,\vec{\delta}^*\right\}$ and α^* that achieves W^*, there exists an equilibrium strategy profile $(\pi^*,\mu^*;\alpha^*)$ such that $W(\pi^*,\mu^*;\alpha^*) = \Phi(\tau^*,\alpha^*) = W^*$. Combining this with (43) establishes the theorem. □

Proof of Proposition 3.

Proof. Note that if $C^*(\tau,\alpha) = 0$ for any $\tau \in \mathcal{D}$ and $\alpha \in \mathcal{A}^*$, the Sender value from persuasion will be his full commitment payoff. Fix any $\tau = \left\{(\beta_j)_{j=1,\dots,|\Omega|};(\delta_j)_{j=1,\dots,|\Omega|}\right\} \in \mathcal{D}$ and $\alpha \in \mathcal{A}^*$. We will show that for each institution, there exists signal $s_1,\dots,s_{|\Omega|} \in \mathcal{S}^B$ and messages $m_1,\dots,m_{|\Omega|} \in M$ such $c(m_j|s_j) = 0$ $\forall j$ and they satisfy constraint (7). For the quadratic and linear distance lying costs, set $s_j = (j-1)\bar{v}$ for $j = 1,2,\dots,|\Omega|$, where, recall that the Sender's payoff is bounded above by $\bar{v}$; and set $m_j = s_j$, so $c(m_j|s_j) = 0$. This implies that $c(m_{j'}|s_j) - c(m_j|s_j) \geq \bar{v}$ $\forall j,j'$, so constraint (7) is always satisfied. Therefore, $C^*(\tau,\alpha) = 0$. Next, for the institution with partial verifiability, if $|\mathcal{V}| \geq |\Omega|$, simply choose any $|\Omega|$ signals from $\mathcal{V}$ and set $m_j = \{s_j\}$ $\forall j$. Next, consider $|\mathcal{V}| = |\Omega| - 1$ and the Sender has state-independent preferences. Without loss, let the beliefs be ordered in the Sender's preference according to index j. For $j \leq |\Omega| - 1$, let s_j be signals in $\mathcal{V}$ and $m_j = s_j$, so constraint (7) are trivially satisfied and $c(m_j|s_j) = 0$. Let $s_{|\Omega|} \in \mathcal{N}$ and $m_{|\Omega|} = s_{|\Omega|}$, so $c\left(m_{|\Omega|}|s_{|\Omega|}\right) = 0$. Given that $m_{|\Omega|}$ induces the best Receiver's belief for the Sender, constraint (7) is also trivially satisfied at $s_{|\Omega|}$. Therefore, $C^*(\tau,\alpha) = 0$ as well. □

Proof of Lemma 2.

Proof. Pick any $|\Omega|$ signals $s_1, s_2, \ldots, s_{|\Omega|}$ from $\mathcal{S}^B$ and let $m_j = s_j\ \forall j$. Since $c\left(m_{j'}|s_j\right) - c\left(m_j|s_j\right) = k$ for any $j \neq j'$, the satisfaction of constraint (8) implies the satisfaction of constraint (7), with $C^*\left(\tau, \alpha\right) = 0$. When constraint (8) is violated, then the feasible sets of signals and messages that satisfy constraint (7) is empty, so $C^*\left(\tau, \alpha\right) = \infty$. □

Proof of Proposition 5.

Proof. By Theorem 1, we have to only consider Bayes plausible distributions of beliefs supported on two beliefs. Let $\underline{\beta}$ and $\bar{\beta}$ denote the two beliefs under the optimal signal structure and, without loss of generality, we assume that $\underline{\beta} \leq \bar{\beta}$. Bayes plausibility implies that $\underline{\beta} \leq \beta^o \leq \bar{\beta}$, and the probabilities of beliefs $\underline{\beta}$ and $\bar{\beta}$ are, respectively, $\underline{\tau} = \frac{\bar{\beta}-\beta^o}{\bar{\beta}-\underline{\beta}}$ and $\bar{\tau} = \frac{\beta^o-\underline{\beta}}{\bar{\beta}-\underline{\beta}}$ when $\underline{\beta} \neq \bar{\beta}$. Since $\hat{v}\left(\beta\right)$ is a singleton $\forall \beta$, we abuse notation and let $\hat{v}\left(\beta\right)$ denote the value of that single element as well. Since $\underline{\beta} \leq \beta^o \leq \frac{L}{1+L}$, $\hat{v}\left(\underline{\beta}\right) = 0$. The Sender's value of persuasion is thus

$$\hat{V}\left(\underline{\beta}, \bar{\beta}\right) = \tau_0 \hat{v}\left(\underline{\beta}\right) + \tau_1 \hat{v}\left(\bar{\beta}\right) = \frac{\beta^o - \underline{\beta}}{\bar{\beta} - \underline{\beta}} \hat{v}\left(\bar{\beta}\right),$$

with Lemma 2 requiring that $\hat{v}\left(\bar{\beta}\right) \leq k$. Since $\bar{\beta} \geq \beta^o$, $\hat{V}\left(\underline{\beta}, \bar{\beta}\right)$ is decreasing in $\underline{\beta}$, so $\underline{\beta} = 0$ and $\hat{V}\left(0, \bar{\beta}\right) = \frac{\beta^o}{\bar{\beta}} \hat{v}\left(\bar{\beta}\right)$. If $\bar{\beta} \leq \frac{L}{1+L}$, $\hat{V}\left(0, \bar{\beta}\right) = 0$; if $\bar{\beta} > \frac{L}{1+L}$, $\hat{V}\left(0, \bar{\beta}\right) = \frac{\beta^o}{\bar{\beta}} \nu\left(\bar{\beta} - L\left(1 - \bar{\beta}\right)\right) > 0$, so $\bar{\beta} > \frac{L}{1+L}$. ν is convex, so it is differentiable almost everywhere, so

$$\begin{aligned}
\frac{d}{d\bar{\beta}} \hat{V}\left(0, \bar{\beta}\right) &\propto \nu'\left(-L + \bar{\beta}\left(1+L\right)\right) - \frac{\nu\left(-L + \bar{\beta}\left(1+L\right)\right)}{\bar{\beta}\left(1+L\right)} \\
&> \nu'\left(-L + \bar{\beta}\left(1+L\right)\right) - \frac{\nu\left(-L + \bar{\beta}\left(1+L\right)\right)}{-L + \bar{\beta}\left(1+L\right)} \geq 0.
\end{aligned}$$

The last inequality follows from the convexity of ν which implies that $\nu'\left(x\right) \geq \frac{\nu(x)}{x}\ \forall x > 0$. Since $\hat{V}\left(0, \bar{\beta}\right)$ is increasing in $\bar{\beta}$, the optimal $\bar{\beta}$ is set such that $\hat{v}\left(\bar{\beta}\right) = k$, thus implying that $\bar{\beta} = \frac{\nu^{-1}(k)+L}{1+L}$ and the Sender's value of persuasion is $V^*\left(k, L\right) = \frac{k\beta^o(1+L)}{\nu^{-1}(k)+L}$. $\frac{d}{dk} V^*\left(k, L\right) = \frac{\beta^o(1+L)}{\left(\nu^{-1}(k)+L\right)^2}\left(L + \nu^{-1}\left(k\right) - k\frac{d\nu^{-1}(k)}{dk}\right)$. ν is convex implies that ν^{-1} is concave, so $\frac{\nu^{-1}(k)}{k} \geq \frac{d\nu^{-1}(k)}{dk}$; therefore, $V^*\left(k, L\right)$ is increasing in k. Finally, $\nu^{-1}\left(k\right) < 1$ implies that $V^*\left(k, L\right)$ is decreasing in L. □

Proof of Lemma 3.

Proof. Fix any $\rho, \rho' \in \Delta\Omega$ and $\tau \in \Delta\Delta\Omega$ such that $\int \sigma d\tau(\sigma) = \rho$.

Euclidean distance: $\psi(\rho|\sigma) = \sqrt{\sum_{\omega\in\Omega} (\sigma(\omega) - \rho(\omega))^2}$**.** Let $||\cdot||$ be the Euclidean norm; therefore, $\psi(\rho|\sigma) = ||\rho - \sigma||$. Recall the reverse triangle inequality $\Big|||x|| - ||y||\Big| \leq ||x - y||$, which implies that

$$\Big|\psi(\rho'|\sigma) - \psi(\rho|\sigma)\Big| \;=\; \Big|||\rho' - \sigma|| - ||\rho - \sigma||\Big| \;\leq\; ||\rho' - \rho||.$$

Therefore,

$$\begin{aligned}\int \psi(\rho'|\sigma) - \psi(\rho|\sigma)\, d\tau(\sigma) \;&\leq\; \int \Big|\psi(\rho'|\sigma) - \psi(\rho|\sigma)\Big| d\tau(\sigma) \\ &\leq ||\rho' - \rho|| \;=\; \psi(\rho'|\rho).\end{aligned}$$

Squared Euclidean distance: $\psi(\rho|\sigma) = \sum_{\omega\in\Omega} (\sigma(\omega) - \rho(\omega))^2$**.**

$$\begin{aligned}\psi(\rho'|\sigma) - \psi(\rho|\sigma) &= \sum_{\omega\in\Omega} (\sigma(\omega) - \rho'(\omega))^2 - \sum_{\omega\in\Omega} (\sigma(\omega) - \rho(\omega))^2 \\ &= \sum_{\omega\in\Omega} \rho'(\omega)^2 - \rho(\omega)^2 - 2\sigma(\omega)\left[\rho'(\omega) - \rho(\omega)\right]\end{aligned}$$

Therefore,

$$\begin{aligned}\int \psi(\rho'|\sigma) - \psi(\rho|\sigma)\, d\tau(\sigma) \;&=\; \sum_{\omega\in\Omega} \rho'(\omega)^2 - \rho(\omega)^2 - 2\int \sigma(\omega)\, d\tau(\sigma)\left[\rho'(\omega) - \rho(\omega)\right] \\ &=\; \sum_{\omega\in\Omega} \rho'(\omega)^2 - \rho(\omega)^2 - 2\rho(\omega)\left[\rho'(\omega) - \rho(\omega)\right] \\ &=\; \sum_{\omega\in\Omega} \left[\rho(\omega) - \rho'(\omega)\right]^2 \;=\; \psi(\rho'|\rho)\end{aligned}$$

Kullback-Leiber divergence: $\psi(\rho|\sigma) = \sum_{\omega\in\Omega} \sigma(\omega) \log \frac{\sigma(\omega)}{\rho(\omega)}$**.**

$$\psi(\rho'|\sigma) - \psi(\rho|\sigma) \;=\; \sum_{\omega\in\Omega} \sigma(\omega) \log \frac{\sigma(\omega)}{\rho'(\omega)} - \sum_{\omega\in\Omega} \sigma(\omega) \log \frac{\sigma(\omega)}{\rho(\omega)} \;=\; \sum_{\omega\in\Omega} \sigma(\omega) \log \frac{\rho(\omega)}{\rho'(\omega)}.$$

Therefore,

$$\begin{aligned}\int \psi\left(\rho'|\sigma\right)-\psi\left(\rho|\sigma\right)d\tau\left(\sigma\right) &= \int\sum_{\omega\in\Omega}\sigma\left(\omega\right)\log\frac{\rho\left(\omega\right)}{\rho'\left(\omega\right)}d\tau\left(\sigma\right)\\ &= \sum_{\omega\in\Omega}\rho\left(\omega\right)\log\frac{\rho\left(\omega\right)}{\rho'\left(\omega\right)} \quad = \quad \psi\left(\rho'|\rho\right)\end{aligned}$$

Symmetric divergence: $\psi\left(\rho|\sigma\right)=\sum_{\omega\in\Omega}\left(\sigma\left(\omega\right)-\rho\left(\omega\right)\right)\log\frac{\sigma(\omega)}{\rho(\omega)}$**.**

$$\begin{aligned}&\psi\left(\rho'|\sigma\right)-\psi\left(\rho|\sigma\right)\\ =&\sum_{\omega\in\Omega}\left[\sigma\left(\omega\right)-\rho'\left(\omega\right)\right]\log\frac{\sigma\left(\omega\right)}{\rho'\left(\omega\right)}-\sum_{\omega\in\Omega}\left[\sigma\left(\omega\right)-\rho\left(\omega\right)\right]\log\frac{\sigma\left(\omega\right)}{\rho\left(\omega\right)}\\ =&\sum_{\omega\in\Omega}\sigma\left(\omega\right)\log\frac{\rho\left(\omega\right)}{\rho'\left(\omega\right)}+\left[\rho\left(\omega\right)-\rho'\left(\omega\right)\right]\log\sigma\left(\omega\right)+\rho'\left(\omega\right)\log\rho'\left(\omega\right)-\rho\left(\omega\right)\log\rho\left(\omega\right)\end{aligned}$$

Therefore,

$$\begin{aligned}&\int\psi\left(\rho'|\sigma\right)-\psi\left(\rho|\sigma\right)d\tau\left(\sigma\right)\\ =&\sum_{\omega\in\Omega}\int\sigma\left(\omega\right)d\tau\left(\sigma\right)\log\frac{\rho\left(\omega\right)}{\rho'\left(\omega\right)}+\left[\rho\left(\omega\right)-\rho'\left(\omega\right)\right]\int\log\sigma\left(\omega\right)d\tau\left(\sigma\right)+\rho'\left(\omega\right)\log\rho'\left(\omega\right)-\rho\left(\omega\right)\log\rho\left(\omega\right)\\ \leq&\sum_{\omega\in\Omega}\int\sigma\left(\omega\right)d\tau\left(\sigma\right)\log\frac{\rho\left(\omega\right)}{\rho'\left(\omega\right)}+\left[\rho\left(\omega\right)-\rho'\left(\omega\right)\right]\log\left(\int\sigma\left(\omega\right)d\tau\left(\sigma\right)\right)+\rho'\left(\omega\right)\log\rho'\left(\omega\right)-\rho\left(\omega\right)\log\rho\left(\omega\right)\\ =&\sum_{\omega\in\Omega}\rho\left(\omega\right)\log\frac{\rho\left(\omega\right)}{\rho'\left(\omega\right)}+\left[\rho\left(\omega\right)-\rho'\left(\omega\right)\right]\log\rho\left(\omega\right)+\rho'\left(\omega\right)\log\rho'\left(\omega\right)-\rho\left(\omega\right)\log\rho\left(\omega\right)\\ =&\sum_{\omega\in\Omega}\left[\rho\left(\omega\right)-\rho'\left(\omega\right)\right]\log\frac{\rho\left(\omega\right)}{\rho'\left(\omega\right)} \quad = \quad \psi\left(\rho'|\rho\right)\end{aligned}$$

□

Proof of Proposition 6.

Proof. Let $(\pi,\mu;\alpha)$ be a Sender-optimal equilibrium. Referring to the proof for Proposition 1, let $\bar{\pi}_2$ be as defined in (24) and $\bar{\mu}_2\left(\cdot|s\right)$ be Dirac on $m=\gamma_2^{-1}\left(s\right)\ \forall s\in\bar{S}_2$. We will show that $\left(\bar{\mu}_2;\alpha\right)$ is also a PBE here with $W\left(\bar{\pi}_2,\bar{\mu}_2;\alpha\right)\geq W\left(\pi,\mu;\alpha\right)$. The arguments with respect to the messaging costs in the proof for Proposition 1 hold here, so we will omit the related details. In particular, (25), (36) and (37) jointly imply that $V\left(\pi,\mu;\alpha\right)=V\left(\bar{\pi}_2,\bar{\mu}_2;\alpha\right)$; therefore, we just have to show that the Sender's best response (which has changed from

(S-IC) to (S-IC')) holds here, and $C(\bar{\pi}_2, \bar{\mu}_2) \leq C(\pi, \mu)$.

By (25), $\rho_{\bar{\pi}_2,\bar{\mu}_2}(\cdot|m) = \rho_{\pi,\mu}(\cdot|m)\ \forall m \in \bar{M}$, so we ease notation as before by letting $\alpha_m(\cdot) = \alpha(\cdot|\rho_{\bar{\pi}_2,\bar{\mu}_2}(\cdot|m)) = \alpha(\cdot|\rho_{\pi,\mu}(\cdot|m))$. Fix $m \in \bar{M}$. For each $s \in \Gamma(m)$, the Sender's best response in (S-IC') holds under (π, μ), hence

$$\begin{aligned} & v\Big(\alpha_m, \sigma_\pi(\cdot|s)\Big) - v\Big(\alpha_{m'}, \sigma_\pi(\cdot|s)\Big) \\ \geq\ & c(m|s) - c(m'|s)\ +\ \psi\Big(\rho_{\pi,\mu}(\cdot|m)\Big|\sigma_\pi(\cdot|s)\Big) - \psi\Big(\rho_{\pi,\mu}(\cdot|m')\Big|\sigma_\pi(\cdot|s)\Big) \quad \forall m' \in M. \end{aligned} \tag{44}$$

As before, let $\Lambda(\cdot|m) \in \Delta S$ be the regular conditional probability measure over S when conditioned on the Receiver receiving message m under (π, μ), which implies that $\rho_{\pi,\mu}(\cdot|m) = \int_{s\in\Gamma(m)} \sigma_\pi(\cdot|s)\, d\Lambda(s|m)$. Recall that $\gamma_2^{-1}(s)$ is the message that the Sender plays with probability one after observing signal s, and $\bar{S}_2$ is the set of signals supported by $\bar{\pi}_2$. Under Assumption 2, $\forall m' \in M$ and $s \in \bar{S}_2$,

$$\begin{aligned} & \psi\Big(\rho_{\bar{\pi}_2,\bar{\mu}_2}(\cdot|m')\Big|\sigma_{\bar{\pi}_2}(\cdot|s)\Big) \\ =& \psi\Big(\rho_{\bar{\pi}_2,\bar{\mu}_2}(\cdot|m')\Big|\rho_{\bar{\pi}_2,\bar{\mu}_2}\left(\cdot|\gamma_2^{-1}(s)\right)\Big) \\ =& \psi\Big(\rho_{\pi_2,\mu_2}(\cdot|m')\Big|\rho_{\pi_2,\mu_2}\left(\cdot|\gamma_2^{-1}(s)\right)\Big) \\ \geq& \int_{s'\in\Gamma\left(\gamma_2^{-1}(s)\right)} \psi\Big(\rho_{\pi_2,\mu_2}(\cdot|m')\Big|\sigma_\pi(\cdot|s')\Big) - \psi\Big(\rho_{\pi_2,\mu_2}\left(\cdot|\gamma_2^{-1}(s)\right)\Big|\sigma_\pi(\cdot|s')\Big) d\Lambda\left(s'|\gamma_2^{-1}(s)\right) \end{aligned} \tag{45}$$

Therefore,

$$\begin{aligned} & v\left(\alpha_{\gamma_2^{-1}(s)}, \sigma_{\bar{\pi}_2}(\cdot|s)\right) - v\left(\alpha_{m'}, \sigma_{\bar{\pi}_2}(\cdot|s)\right) \\ =& \int_{s'\in\Gamma\left(\gamma_2^{-1}(s)\right)} \left[v\left(\alpha_{\gamma_2^{-1}(s)}, \sigma_\pi(\cdot|s')\right) - v\left(\alpha_{m'}, \sigma_{\pi_2}(\cdot|s')\right)\right] d\Lambda\left(s'|\gamma_2^{-1}(s)\right) \qquad (46) \\ \geq& \int_{s'\in\Gamma\left(\gamma_2^{-1}(s)\right)} \left[c\left(\gamma_2^{-1}(s)|s'\right) - c(m'|s')\right] d\Lambda\left(s'|\gamma_2^{-1}(s)\right) \\ & + \int_{s'\in\Gamma\left(\gamma_2^{-1}(s)\right)} \left[\psi\Big(\rho_{\pi,\mu}\left(\cdot|\gamma_2^{-1}(s)\right)\Big|\sigma_\pi(\cdot|s')\Big) - \psi\Big(\rho_{\pi,\mu}(\cdot|m')\Big|\sigma_\pi(\cdot|s')\Big)\right] d\Lambda\left(s'|\gamma_2^{-1}(s)\right) \\ & \qquad (47) \\ \geq& c\left(\gamma_2^{-1}(s)|s\right) - c(m'|s) + \underbrace{\psi\Big(\rho_{\bar{\pi}_2,\bar{\mu}_2}\left(\cdot|\gamma_2^{-1}(s)\right)\Big|\sigma_{\bar{\pi}_2}(\cdot|s)\Big)}_{=0} - \psi\Big(\rho_{\bar{\pi}_2,\bar{\mu}_2}(\cdot|m')\Big|\sigma_{\bar{\pi}_2}(\cdot|s)\Big) \qquad (48) \end{aligned}$$

This implies the Sender's best response; therefore, $(\bar{\mu}_2; \alpha)$ is a PBE under $\bar{\pi}_2$. Next,

$$\begin{aligned}
C\left(\bar{\pi}_2, \bar{\mu}_2\right) =& \sum_{\omega' \in \Omega} \beta^o\left(\omega'\right) \int_{s \in S} \int_{m \in \bar{M}} \left[c\left(m|s\right) + \psi\Big(\rho_{\bar{\pi}_2, \bar{\mu}_2}\left(\cdot|m\right) \Big| \sigma_{\bar{\pi}_2}\left(\cdot|s\right)\Big)\right] d\bar{\mu}_2\left(m|s\right) d\bar{\pi}_2\left(s|\omega'\right) \\
=& \sum_{\omega' \in \Omega} \beta^o\left(\omega'\right) \int_{s \in S} \int_{m \in \bar{M}} c\left(m|s\right) d\bar{\mu}_2\left(m|s\right) d\bar{\pi}_2\left(s|\omega'\right) \\
\leq& \sum_{\omega \in \Omega} \beta^o\left(\omega\right) \int_{s \in S} \int_{m \in \bar{M}} c\left(m|s\right) d\mu\left(m|s\right) d\pi\left(s|\omega\right) \qquad (49) \\
\leq& \sum_{\omega \in \Omega} \beta^o\left(\omega\right) \int_{s \in S} \int_{m \in \bar{M}} \left[c\left(m|s\right) + \psi\Big(\rho_{\pi, \mu}\left(\cdot|m\right) \Big| \sigma_{\pi}\left(\cdot|s\right)\Big)\right] d\mu\left(m|s\right) d\pi\left(s|\omega\right) \\
=& C\left(\pi, \mu\right)
\end{aligned}$$

Therefore, $W\left(\bar{\pi}_2, \bar{\mu}_2; \alpha\right) \geq W\left(\pi, \mu; \alpha\right)$, so $W\left(\bar{\pi}_2, \bar{\mu}_2; \alpha\right)$ is also a Sender-optimal equilibrium. The last part of the proposition follows from the same argument as Proposition 2. □