Strategic Interpretations[*]

Kfir Eliaz[†], Ran Spiegler[‡] and Heidi C. Thysen[§]

October 29, 2020

Abstract

We study strategic communication when the sender's multi-dimensional messages are given an interpretation by the sender himself or by a proxy. Interpreting messages involves the provision of some data about their statistical state-dependence. Interpretation can be selective: different kinds of data interpret different sets of message components. The receiver uses this data to decipher messages, yet he does not draw any inferences from the kind of data he is given. In this way, strategic interpretation of messages can influence the receiver's understanding of their equilibrium meaning. We show that in a two-action, two-state setting, the sender can attain his first-best payoff when the prior on one state exceeds a threshold that decays quickly with message dimensionality. We examine the result's robustness to the critique that our receiver does not attempt any inferences from selective interpretations.

[*] Financial support by ERC Advanced Investigator grant no. 692995 is gratefully acknowledged. Thysen gratefully acknowledges financial support by ESRC award no. 1779091. We thank Xiaosheng Mu, Ariel Rubinstein, Joel Sobel and seminar audiences at Columbia, Hebrew and Penn universities as well as the Warwick Economic Theory conference, for helpful comments.

[†] School of Economics, Tel-Aviv University and David Eccles School of Business, the University of Utah. E-mail: kfire@post.tau.ac.il.

[‡] School of Economics, Tel-Aviv University and Economics Dept., University College London and CFM. E-mail: rani@tauex.tau.ac.il.

[§] London School of Economics. E-mail: h.c.thysen@lse.ac.uk.

1 Introduction

In the simplest textbook model of strategic communication, originated by Crawford and Sobel (1982), a "sender" privately observes a state of Nature and chooses a costless message from some given message space. Then, a "receiver" observes the message and takes an action that affects both parties' payoffs. A hallmark of this conventional approach is that messages have no intrinsic meaning; their content - namely, their statistical relation with the underlying state - is established in Nash equilibrium of the sender-receiver game. According to the standard steady-state interpretation of this solution concept, the receiver has access to a "dataset" that fully reveals the statistical relation between states and messages.

In this paper we revisit the basic sender-receiver model and relax the assumption that the receiver is fully capable of interpreting equilibrium messages. We focus on settings in which the receiver has two available actions, y and n. In each state of Nature, exactly one of these actions is appropriate. The prior probability of the states for which y is the appropriate action is $\pi < \frac{1}{2}$. The receiver's sole objective is to select the appropriate action. This familiar setting is borrowed from Glazer and Rubinstein (2004, 2006) or Kamenica and Gentzkow (2011). For most of the paper, we follow these papers by also assuming that the sender always wants the receiver to play y (but we also examine an alternative, "zero-sum" specification).

By default, our receiver lacks access to any data regarding the state-message mapping, and therefore cannot decipher messages by himself. He is like a tourist in a foreign country who does not understand its language or cultural codes. However, if an "interpreter" handed him a "*dictionary*" containing data regarding the statistical mapping from states to the sender's messages, he would have some ability to interpret the message he receives.

Our model makes room for the *strategic* supply of such dictionaries. The sender himself - or a *third party* who acts as an *interpreter* on the sender's behalf - chooses a dictionary from some feasible set. He can condition the dictionary on the state and the message. Thus, different messages may be accompanied by different dictionaries, and the same message may be paired with different dictionaries in different states.

Each dictionary provides *credible*, yet possibly selective statistical data regarding the sender's state-message mapping (given by the sender's strategy). The receiver uses this data to update his belief given the message. Crucially, our basic model assumes that the receiver lacks any other means for extracting the meaning of messages (we relax this assumption in Section 4). Consequently, he *does not draw any inferences from the provided dictionary itself*, since this would require some data regarding the *joint* distribution of messages, dictionaries and states - data the receiver does not have.[1] Consequently, the sender can manipulate the receiver's beliefs beyond what is feasible under rational expectations.

Strategic interpretation of messages - in the sense of providing selective statistical data about their meaning - is pervasive in real-life situations, whether the messages are cheap talk or hard-information disclosures. Consider an employee who wants to exert effort only when sufficiently sure he is not about to be fired. He is summoned to the General Manager's office to hear about his prospects at the company. After the meeting is over, the HR manager (who was present at the meeting) explains that when the GM says to an employee "you have a future in the company", this means a 50% chance of keeping his job. This is an interpretation of the GM's verbal message. It is selective because it ignores other aspects of the GM's communication, e.g. his body language. Alternatively, the HR's interpretation could focus on the latter: "The GM's handshake was feeble; this is definitely bad news".

Another example involves a tenure case that is brought in front of a university promotions committee. Although the candidate submits his CV, committee members outside his discipline cannot decipher the connection between the candidate's quality and indicators such as the number of publications, conference lectures or supervised students. The candidate's department chair will offer an interpretation by providing statistical data about researchers in comparable departments (including their subsequent academic performance, which indicates their "true quality"). If the chair's objective is misaligned with the university committee's, the data he provides may be strategically selective. In a similar vein, imagine a foreign candidate for a graduate

[1] A similar form of bounded rationality is documented in Jin et al. (2019), who find that in a laboratory game of voluntary disclosure, receivers do not make correct inferences from no disclosure.

program. The candidate submits his grade transcript, yet the admission committee does not know the grades' meaning. A faculty member writing a recommendation letter on the candidate's behalf may provide such an interpretation, by describing the grade distribution for a selected subset of courses.

Finally, suppose the sender is a political party and the receiver is a representative voter. The party's message is multi-dimensional, where each component describes public pronouncements by a different party member. A political commentator interprets the party's message in some media outlet. He does so by providing historical data about the match between the public pronouncements of selected party organs and the underlying reality.

These are all examples of selective interpretations where the receiver is presented with partial statistics about the sender's state-dependent, multi-dimensional message. These interpretations can be strategic when the interpreter's interests are misaligned with the receiver's. We analyze the sender's choice of messages when he takes their subsequent strategic interpretation into account. For instance, the way a political party structures the public statements made by its members will be shaped by its expectation of how a media outlet that is biased in its favor will interpret these statements.

One could argue that in these examples, the statistical data the interpreter provides need not be perfectly credible or unbiased. However, because they are quantitative and verifiable, they are more likely to be credible than cheap-talk messages like "you have a future in the company". At any rate, we abstract from this consideration; our analytical task is to quantify the effect of strategic provision of *cheap-talk* messages and their interpretation on the sender's ability to attain his objective, assuming perfect credibility of the statistical data these interpretations involve. In the course of this paper, we will consider various kinds of partial statistics that strategic interpretations can entail.

Preview of the analysis

We present our basic model in Section 2, where we define a dictionary as a non-empty subset of the components of a K-dimensional message. The dictionary enables the receiver to learn the state-dependent joint distribution of these components. We assume that the interpreter's preferences fully coincide with the sender's. For expositional convenience, our formal exposition regards them as a *single* player who *commits* to a state-dependent joint distribution over messages and dictionaries. Neither of these two assumptions is necessary for our main findings. (In our informal description, we occasionally refer to the interpreter as a distinct agent who shares the sender's preferences.)

In Section 3, we present our main result, which characterizes the maximal probability of persuasion as a function of π and K. In particular, we show that the sender can attain full persuasion, as long as π is above a cutoff $\pi^*(K)$ given by a simple formula that makes use of Sperner's Theorem and decays quickly with K.

Our assumption that the receiver draws no inferences from the dictionary he is given raises natural questions. First, does the dictionary itself convey information about the underlying state? The answer is negative: The sender-optimal strategy we construct has the property that the distribution over dictionaries is state-independent. Second, would the receiver be "suspicious" of a dictionary that does not cover all message components? We address this question in Section 4, while insisting on sender strategies that induce a state-independent dictionary distribution.

In Section 4.1, we perturb the model by assuming that the sender has a lexicographically secondary preference for small dictionaries. We also introduce a refinement of the sender's strategy: if the sender's interests were aligned with the receiver's, he would want to play a strategy that induces the same observed distribution over dictionaries. Thus, if the receiver had independent access to data about the distribution of dictionaries, he could reconcile the observed use of selective dictionaries with a benevolent sender. Under this refinement, we show that full persuasion is attainable if and only if $\pi \geq 1/(K+1)$. The sender's strategy only interprets *single* message components.

In Section 4.2, we modify the definition of dictionaries. When a dictionary $D \subseteq \{1,...,K\}$ is provided, this now means that the receiver learns the state-dependent distribution of m_D *as well as* the state-dependent distribution of $m_{\{1,...,K\}\setminus D}$. Thus, the interpreter is forced to provide statistical data about the behavior of *all* message components, though in a format that can break them into two disjoint sets. Under a mild assumption on how the receiver extrapolates a belief from these pieces of data, we show that full persuasion is attainable whenever π exceeds a cutoff that decays quickly with K. The lesson from these two variants of our basic model is that strategic interpretation can produce effective persuasion without generating excessive "suspicion" regarding its selectivity.

Section 5 picks up the theme of Section 4.2 and present an example that illustrates a richer notion of dictionaries, which involves data about other slices of the joint state-message distribution. We show how this richer specification can enhance the sender's ability to attain full persuasion. In Section 6 we perform partial analysis of our basic model when the two parties have diametrically opposed preferences. We discuss related literature in Section 7.

2 A Model

There are two players, a sender and a receiver. The sender observes a state of Nature $\theta \in \Theta = \{Y, N\}$. The receiver does not observe the state but needs to take an action a, which can be either "yes" (denoted y) or "no" (denoted n). Players' payoffs take values in $\{0, 1\}$. The receiver's payoff is 1 if either $< a = y$ and $\theta = Y >$ or $< a = n$ and $\theta = N >$, and 0 otherwise. In contrast, the sender's payoff is 1 if and only if $a = y$, and 0 otherwise.

The players' common prior belief over Θ assigns probability $\pi < \frac{1}{2}$ to state Y. Hence, the receiver's ex-ante optimal action is n. However, the sender can influence the receiver's belief and persuade him to play y. He commits to a strategy that maps each state to a distribution over *reports,* where a report is a pair (m, D) such that:

(i) $m = (m_1, ..., m_K) \in M^K$ is a K-dimensional *message*, where $K \geq 1$ and $|M| \geq 2$. In all the examples we use in the paper, $M = \{0, 1\}$.

(ii) $D \in 2^{\{1,...,K\}} \backslash \{\emptyset\}$ is a *dictionary.*

Thus, the sender's strategy is a function $\sigma : \Theta \to \Delta\left(M^K \times 2^{\{1,...,K\}} \backslash \{\emptyset\}\right)$. The commitment assumption is made for expositional simplicity; as we shall see, our results regarding full persuasion are insensitive to it. The assumption that $|\Theta| = 2$

could be replaced with the weaker assumption that there is a function $f : \Theta \rightarrow \{n, y\}$ such that the receiver's payoff is 1 if and only if $a = f(\theta)$, and 0 otherwise. The probability with which the sender plays the report (m, D) in state θ is denoted $\sigma(m, D \mid \theta)$. With slight abuse of notation, define $\sigma(m \mid \theta) = \sum_D \sigma(m, D \mid \theta)$ and $\sigma(D \mid \theta) = \sum_m \sigma(m, D \mid \theta)$. We refer to $(\sigma(m \mid \theta))$ as the *message strategy* and to $(\sigma(D \mid m, \theta))$ as the *interpretation strategy.*

The role of dictionaries is to grant the receiver "partial access" to the statistical regularities of the sender's strategy. When the receiver observes the report (m, D), he learns the conditional probabilities $(\sigma(m_D \mid \theta))_{\theta \in \Theta}$, where $m_D = (m_k)_{k \in D}$ and

$$\sigma(m_D \mid \theta) = \sum_{m' \mid m'_D = m_D} \sigma(m' \mid \theta)$$

That is, the receiver learns how the message components in D - *and nothing but them* - are distributed conditional on the state. He *cannot* draw any statistical inferences from the message components $m_{\{1,...,K\} \setminus D}$ or the dictionary D itself. We will revisit this assumption in the sequel. Note that in any report (m, D), D must be a *non-empty* subset of $\{1, ..., K\}$; that is, the sender is obliged to provide *some* interpretation of the message.

Upon receiving a report (m, D), the receiver updates his belief according to the following expression:

$$\widetilde{\Pr}(\theta = Y \mid m, D) = \frac{\pi \cdot \sigma(m_D \mid \theta = Y)}{\pi \cdot \sigma(m_D \mid \theta = Y) + (1 - \pi) \cdot \sigma(m_D \mid \theta = N)} \tag{1}$$

Compare this with the correct, rational-expectations posterior probability of Y conditional on (m, D):

$$\Pr(\theta = Y \mid m, D) = \frac{\pi \cdot \sigma(m, D \mid \theta = Y)}{\pi \cdot \sigma(m, D \mid \theta = Y) + (1 - \pi) \cdot \sigma(m, D \mid \theta = N)} \tag{2}$$

The receiver best-replies to the subjective posterior belief given by (1), breaking ties in favor of the sender. Equivalently, faced with a report (m, D), he computes its subjective likelihood ratio

$$\rho_\sigma(m, D) = \frac{\sum_{m' | m'_D = m_D} \sigma(m' \mid \theta = Y)}{\sum_{m' | m'_D = m_D} \sigma(m' \mid \theta = N)} \tag{3}$$

and chooses $a = y$ if and only if $\rho_\sigma(m, D) \geq (1 - \pi)/\pi$.

The sender chooses his strategy under the assumption that the receiver best-replies to the belief given by (1). Our main question is: What is the maximal probability of $a = y$ that the sender can attain?

Our model of how the receiver forms beliefs is motivated by the steady-state view of equilibrium behavior, whereby the sender's strategy σ describes a long-run statistical relation between states and reports. The receiver moves once, against the background of a large dataset consisting of many realizations of $(\theta, m_1, ..., m_K, D)$ resulting from previous interactions between the sender with different identical receivers. The dataset can be visualized as a large spreadsheet, where each column represents one of the variables $\theta, m_1, ..., m_K, D$, and each row represents an observation (an independent draw from the joint distribution over states and reports). Rational expectations correspond to having full access to this dataset. Our model relaxes this assumption and assumes that the receiver is granted access to a subset of columns represented by D. The receiver can only rely on the accessed data for drawing inferences.

Example 1

To illustrate our notion of dictionaries and how the receiver reacts to them, suppose that $K = 4$. Assume $\sigma(m \mid Y)$ is uniform over $(1, 1, 1, 1)$ and $(0, 0, 0, 0)$, while $\sigma(m \mid N)$ is uniform over $(1, 1, 1, 1)$, $(1, 0, 1, 0)$ and $(1, 0, 0, 1)$.

Suppose the sender accompanies the message $(1, 0, 1, 0)$ with the dictionary $D = \{1, 3\}$. This dictionary provides the receiver with data about the state-dependent distribution of (m_1, m_3). In particular, he learns that the pattern $(1, *, 1, *)$ occurs with probability $\frac{1}{2}$ in state Y and with probability $\frac{2}{3}$ in state N.[2] Therefore,

[2] The notation $(1, *, 1, *)$ stands for all messages m for which $m_1 = m_3 = 1$.

$$\widetilde{\Pr}(\theta = Y \mid (1,0,1,0), \{1,3\}) = \frac{\pi \cdot \frac{1}{2}}{\pi \cdot \frac{1}{2} + (1-\pi) \cdot \frac{2}{3}} = \frac{3\pi}{4-\pi}$$

By comparison, the rational-expectations posterior on Y given $m = (1,0,1,0)$ is zero (independently of the dictionary that accompanies this message).

Note that the message $(1,1,1,1)$ is sent with positive probability in *both* states. Suppose that in state Y the sender accompanies this message with the dictionary $D = \{1,2,3\}$. The receiver then learns that the pattern $(1,1,1,*)$ occurs with probability $\frac{1}{2}$ in state Y and with probability $\frac{1}{3}$ in state N. Hence,

$$\widetilde{\Pr}(\theta = Y \mid (1,1,1,1), \{1,2,3\}) = \frac{\pi \cdot \frac{1}{2}}{\pi \cdot \frac{1}{2} + (1-\pi) \cdot \frac{1}{3}} = \frac{3\pi}{2+\pi}$$

Suppose next that in state N the sender accompanies the message $(1,1,1,1)$ with the dictionary $D = \{3\}$. Then

$$\widetilde{\Pr}(\theta = Y \mid (1,1,1,1), \{3\}) = \frac{\pi \cdot \frac{1}{2}}{\pi \cdot \frac{1}{2} + (1-\pi) \cdot \frac{2}{3}} = \frac{3\pi}{4-\pi}$$

Thus, by varying the dictionary across states, the same message induces the receiver to hold a different belief in each state. In contrast, if the receiver had rational expectations, then independently of the dictionary, his posterior on Y given $m = (1,1,1,1)$ would be $\frac{3\pi}{2+\pi}$ in *both* states. □

We close this section with comments on a few aspects of our model.

Multi-dimensional messages

The multi-dimensionality of messages has a few interpretations. First, different components of m may represent different modes of communication (verbal statements, voice intonation). When the sender is an organization, different components represent utterances by different organs (party whip, corporate executive, spokesperson). Finally, the state itself can be multi-dimensional (this requires $|\Theta| > 2$), such that each message component corresponds to a different state dimension.

Rational expectations and the full dictionary

Note that the full dictionary $D = \{1, ..., K\}$ does *not* automatically endow the receiver with rational expectations. The reason is that rational expectations mean that the receiver knows the sender's entire reporting strategy, whereas the full dictionary only enables him to learn the message strategy. However, if the interpretation strategy happens to be measurable with respect to messages (i.e. $\sigma(D \mid m) \equiv \sigma(D \mid m, \theta)$), accompanying a message with the full dictionary will enable the receiver to update his belief as if he had rational expectations.

The "redacted message" metaphor

Our model could be alternatively described as follows. When the sender sends a message, he selectively "redacts" parts of that message, such that the receiver gets to observe only the unredacted parts. The belief-formation rule (1) means that the receiver takes into account the sender's pre-redaction message strategy but ignores the redaction strategy (and therefore draws no inference from the redacted components).

We find this "selective redaction" description less appealing because it lacks a concrete story for how the receiver forms correct expectations about the sender's message strategy but not about the redaction strategy. In contrast, our original description of D as a representation of selective statistical data regarding the sender's strategy entails an explicit mechanism for this dichotomy: The receiver can only base his beliefs on the statistical data provided to him by the sender.

More importantly, our description opens the door for *other types of dictionaries* that correspond to other kinds of statistical data that the sender can transmit to the receiver. We illustrate this idea in Sections 4.2 and 5, where we allow the sender to provide multiple "datasets" that record different slices of the joint state-message distribution, and show how this richer notion of dictionaries affects the sender's problem. These extensions of our basic model go beyond the scope of the "redaction" metaphor.

Who interprets the messages?
Given that we model the situation as a two-player game, a literal interpretation of our model would be that the sender interprets his own messages. A more plausible story is that the two-player model is a reduced form of a *larger* model, in which interpretation is done by a *third party* whose preferences are aligned with the sender's: An accomplice, a spokesperson or a captured media outlet. Such third parties provide selective data that illuminate the meaning of utterances by the agent they serve.

We could turn the interpreter into an actual third player, producing the following timeline. The sender moves first by choosing a message m. The interpreter moves after observing m (but not θ) and chooses D. This means that the interpretation strategy must be measurable with respect to m. Unlike the receiver, the interpreter has rational expectations. The conditional distribution σ over pairs (m, D) is induced by the combination of the message and interpretation strategies, and it is restricted to satisfy the conditional-independence property $D \perp \theta \mid m$. The receiver moves last, having observed the history (m, D), and he best-replies to the belief (1). If the sender and interpreter have common interests, the situation can be reduced to our two-player formulation, under a suitably defined solution concept for the three-player interaction. In Section 3 we will see that there is no loss of generality in imposing $D \perp \theta \mid m$ directly on the two-player model, lending support to this three-player interpretation of our model.

Thus, while we will adhere to the sender-receiver formal terminology, our model can be regarded as a description of a situation in which the sender and interpreter are separate entities who happen to share common interests.[3]

3 Analysis

We begin this section by presenting the rational-expectations benchmark for our model. In this case, which coincides with the "prosecutor" example in Kamenica

[3] In a previous version of the paper (Eliaz et al., 2018), we analyzed an extension in which the interpreter's preferences are aligned with the receiver's with some probability; the sender does not know the interpreter's type when choosing his message strategy.

and Gentzkow (2011), the probability of persuasion is maximized by the following message strategy (the dictionary component in the reporting strategy is redundant): In state Y, the sender plays $(1, ..., 1)$ with probability one, whereas in state N, he plays $(1, ..., 1)$ with probability $\pi/(1-\pi)$ and $(0, ..., 0)$ with the remaining probability. When the receiver gets the message $(0, ..., 0)$, he infers that $\theta = N$ for sure and takes the action n. When he receives the message $(1, ..., 1)$, his posterior is

$$\Pr(\theta = Y \mid m = (1, ..., 1)) = \frac{\pi \cdot 1}{\pi \cdot 1 + (1-\pi) \cdot \frac{\pi}{1-\pi}} = \frac{1}{2}$$

such that he is just willing to play y. Consequently, the overall probability of persuasion is

$$\pi + (1-\pi) \cdot \frac{\pi}{1-\pi} = 2\pi$$

This result crucially relies on the sender's ability to *commit* to a strategy ex-ante. Without the ability to commit, the probability of persuasion would be *zero* in any Nash equilibrium.

The following example demonstrates that in contrast to the rational-expectations benchmark, our model enables *full* persuasion as an equilibrium outcome.

Example 2: Full persuasion under $K = 3$ *and* $K = 4$
Let $K = 3$ and consider the following sender strategy (for convenience, we highlight the interpreted components in each report in boldface). In each state, he mixes uniformly over three reports:

State Y		State N	
m	D	m	D
111	$\{1\}$	**1**00	$\{1\}$
1**1**1	$\{2\}$	0**1**0	$\{2\}$
11**1**	$\{3\}$	00**1**	$\{3\}$

Notice that in state Y only one message is sent, but the sender randomizes the dictionary it is paired with. In contrast, in state N, three distinct messages are sent with three distinct dictionaries, where each dictionary interprets a pattern that also appears in state Y (namely, the component with the digit 1). For each of the six reports $(m, \{k\})$, the receiver's posterior belief $\widetilde{\Pr}(\theta = Y \mid m, \{k\})$ is

$$\frac{\pi \cdot \Pr(m_k = 1 \mid \theta = Y)}{\pi \cdot \Pr(m_k = 1 \mid \theta = Y) + (1-\pi) \cdot \Pr(m_k = 1 \mid \theta = N)} = \frac{\pi \cdot 1}{\pi \cdot 1 + (1-\pi) \cdot \frac{1}{3}} = \frac{3\pi}{1+2\pi}$$

The receiver weakly prefers playing y after each of these reports, as long as $\pi \geq \frac{1}{4}$.

If $K = 4$, the sender is able to achieve full persuasion for even smaller prior beliefs. He achieves this by using the following strategy, which in each state, uniformly randomizes over six reports:

State Y		State N	
m	D	m	D
1111	$\{1,2\}$	**11**00	$\{1,2\}$
11**1**1	$\{1,3\}$	**1**0**1**0	$\{1,3\}$
111**1**	$\{1,4\}$	**1**00**1**	$\{1,4\}$
1**11**1	$\{2,3\}$	0**11**0	$\{2,3\}$
1**1**1**1**	$\{2,4\}$	0**1**0**1**	$\{2,4\}$
11**11**	$\{3,4\}$	00**11**	$\{3,4\}$

For each of these twelve reports $(m, \{j, k\})$, the receiver's posterior belief $\widetilde{\Pr}(\theta = Y \mid m, \{j, k\})$ is

$$\frac{\pi \cdot \Pr(m_j = m_k = 1 \mid \theta = Y)}{\pi \cdot \Pr(m_j = m_k = 1 \mid \theta = Y) + (1-\pi) \cdot \Pr(m_j = m_k = 1 \mid \theta = N)} = \frac{\pi \cdot 1}{\pi \cdot 1 + (1-\pi) \cdot \frac{1}{6}}$$

The receiver weakly prefers playing y after each of these reports, as long as $\pi \geq \frac{1}{7}$.

This example illustrates a number of key points.

Non-rational expectations

The receiver reaches wrong beliefs as a result of the strategically chosen dictionaries. E.g., in the $K = 4$ case, although the reports $((1,1,1,1), \{2,3\})$ and $((0,1,1,0), \{2,3\})$

objectively reveal the state in which they are played, the receiver draws the same inference from both of them. The reason is that the two messages coincide on the second and third components, highlighted by the accompanying dictionary $\{2, 3\}$.

Irrelevance of commitment
Since the sender achieves full persuasion, his strategy would also constitute an equilibrium in the *absence* of commitment. The reason is that the receiver plays y after any realized report, hence the sender has no incentive to deviate from any realization of his mixed strategy.

More (interpretation) can be less
The receiver is clearly harmed by selective interpretation: If the sender were compelled to interpret all message components, the problem would be effectively reduced to the rational-expectations benchmark. However, this effect is not monotone. Suppose that we made dictionaries even *more* selective by forcing them to be *singletons*. Then, in the $K = 4$ case, the sender would only be able to attain full persuasion when $\pi \geq \frac{1}{5}$, using a similar strategy to the one we presented for $K = 3$.

Dictionary-state independence
Our model assumes that the receiver cannot draw any inferences from D. Suppose he attempted such an inference - e.g. by acquiring data regarding the state-contingent distribution over dictionaries. Then, he would be unable to infer the state from D because its probability is identical in both states. One might argue that the receiver should still be "suspicious" of selective interpretations and discount their informational content. We devote Section 4 to this critique.

The sender's strategy satisfies another independence property: $D \perp \theta \mid m$. That is, given the realized message, the dictionary that accompanies it does not provide objective information about the state. This means that if the receiver had rational expectations, he could afford to draw inferences from m alone. The following lemma establishes that this property is not specific to the example.

Lemma 1 *The maximal probability of persuasion can be attained by a strategy that satisfies $D \perp \theta \mid m$.*

Proof. Consider an arbitrary sender strategy σ. Suppose that for a given message m there are two dictionaries D and D', such that both reports (m, D) and (m, D') are played with positive probability under σ. Suppose without loss of generality that the action induced by (m, D) is weakly more favorable to the sender (recall that the sender's preferences are state-independent). Consider a deviation that replaces (m, D') with (m, D). Since the deviation does not change the message strategy, it does not affect the receiver's reaction to any report $(m'', D'') \neq (m, D)$; and by increasing the probability of (m, D), it weakly increases the probability of persuasion. It follows that without loss of generality, we can assume that under the sender's strategy, every realized report m is accompanied by a *single* dictionary D_m. In particular, this means that D is independent of θ conditional on m. ■

This lemma substantiates the three-player interpretation of our model that was described at the end of Section 2, since a distinct interpreter would only be able to condition D on m.

Should a dictionary interpret multiple messages?

The sender's strategy in Example 2 has a notable feature: In every report (m, D) that is played in state N, the pattern that D highlights does not appear in any other message that is played in N. Compare this with the report $((1, 0, 1, 0), \{1, 3\})$ in Example 1. The dictionary $\{1, 3\}$ highlights the pattern $(1, *, 1, *)$, which also appears in *another* message, $(1, 1, 1, 1)$, that is played in state N. It turns out that this feature of the sender's behavior in Example 1 is weakly sub-optimal. That is, when solving the sender's problem, we can restrict attention to strategies that satisfy the following property: for every persuasive report (m, D) that is played in state N, D highlights a pattern that does not appear in any other message sent in that state. This property will facilitate the proof of our main result.

Fix a sender's strategy σ. Let $\mathcal{B}_\sigma$ be the set of reports (m, D) that are played with positive probability in $\theta = N$ and persuade the receiver. That is,

$$\mathcal{B}_\sigma = \left\{ (m, D) \mid \sigma(m, D \mid \theta = N) > 0 \text{ and } \rho_\sigma(m, D) \geq \frac{1 - \pi}{\pi} \right\}$$

Proposition 1 *For every sender strategy σ, there exists a strategy σ' with the following properties:* (*i*) *the probability that the receiver chooses y in each state is at least as high as under σ, and* (*ii*) $m'_D \neq m_D$ *for every pair of distinct reports* $(m, D), (m', D') \in \mathcal{B}_{\sigma'}$.

Our proof employs a two-stage algorithm. In the first stage, we replace "redundant dictionaries". We list the reports in $\mathcal{B}_\sigma$ according to an arbitrary ordering. Then, starting with the report (m, D) at the top of the ordering, we identify messages m' down the list such that $m'_D = m_D$. We then replace the dictionaries that accompany these messages with D. Setting aside the top report and all the reports that were subjected to this replacement, we continue in the same manner with the remaining reports. At the end of the algorithm's first stage, $\mathcal{B}_\sigma$ is partitioned such that each cell consists of reports (m, D) with the same D and m_D. In the second stage, we replace "redundant messages". We go *up* the list of reports and modify messages only, such that each cell in the above partition ends up consisting of a single report. (We may perform additional changes to the dictionaries that accompany messages in state Y, to ensure that probability that $a = y$ in this state does not go down.)

Our subsequent analysis makes use of the following concept.

Definition 1 *For a given a strategy σ, a message m' is said to* ***justify*** *the report* $(m, D) \in \mathcal{B}_\sigma$ *if:* (*i*) $\sigma(m' \mid \theta = Y) > 0$, *and* (*ii*) $m'_D = m_D$.

In other words, what helps persuade the receiver to choose y when he gets the report (m, D) is that the pattern highlighted by D appears in some messages m' that are played with sufficient frequency in state Y.

Proposition 1 is particularly useful because it places restrictions on the family of reports that any given message can justify. This is captured by the following corollaries.

Corollary 1 *Let* $(m, D), (m', D') \in \mathcal{B}_\sigma$. *If there is a message m^* that justifies both* (m, D) *and* (m', D'), *then* $D \nsubseteq D'$ *and* $D' \nsubseteq D$.

Corollary 2 *The number of reports that any message justifies is at most* $\binom{K}{\lfloor K/2 \rfloor}$.

Corollary 1 says that the set of dictionaries that appear in reports that are justified by a given message m^* constitutes an *anti-chain* - i.e., no dictionary in this set contains another. Corollary 2 then invokes Sperner's Theorem. This fundamental result in extremal combinatorics states that the largest anti-chain over $\{1, 2, \ldots, K\}$ is the collection of all subsets of size $\lfloor K/2 \rfloor$.

We are now ready to state the main result of this section. The result makes use of the following notation, which will also serve us in later sections:

$$\begin{aligned} S &= \binom{K}{\lfloor K/2 \rfloor} \\ \mathcal{B}^* &= \left\{ (m, D) \mid m_k = \mathbf{1}(k \in D) \ ; \ |D| = \left\lfloor \frac{K}{2} \right\rfloor \right\} \end{aligned}$$

Note that $|\mathcal{B}^*| = S$.

Theorem 1 *The maximal probability of persuasion is* $\min\{1, \pi(1+S)\}$. *It can be implemented by the following strategy:*

$$\begin{aligned} \sigma((1, \ldots, 1), D \mid \theta = Y) &= \frac{1}{S} \textit{ for every } D \textit{ for which } |D| = \left\lfloor \frac{K}{2} \right\rfloor \\ \sigma(m, D \mid \theta = N) &= \min\{\frac{1}{S}, \frac{\pi}{1-\pi}\} \textit{ for every } (m, D) \in \mathcal{B}^* \\ \sigma((0, \ldots, 0), D \mid \theta = N) &= \max\{0, \frac{1}{S} - \frac{\pi}{1-\pi}\} \textit{ for every } D \textit{ for which } |D| = \left\lfloor \frac{K}{2} \right\rfloor \end{aligned}$$

Furthermore, when $\pi \geq 1/(1+S)$, *this strategy is time-consistent and attains full persuasion.*

The strategy that implements the maximal probability of persuasion generalizes Example 2. In state Y, the sender sends a single message, which we conveniently select to be $(1, ..., 1)$. Each of the components of this message can therefore be regarded as "good news". What happens in state N depends on the relation between the prior π and the number S, which depends on K. Suppose K is even, for the sake of the argument. If $\pi \geq 1/(1+S)$, the sender randomizes uniformly over $\mathcal{B}^*$, which is the set of all reports in which the message consists of an equal number of 1's ("good news") and 0's ("bad news"), and the dictionary interprets only the good news. If $\pi < 1/(1+S)$, each of these reports is played with probability $1/S$, and the remaining probability is allocated to the message $(0, ..., 0)$ - i.e. all "bad news".

Unlike the case of the "mixed" messages in $\mathcal{B}^*$, there is considerable freedom in selecting the dictionaries that accompany the "pure" messages $(1, ..., 1)$ and $(0, ..., 0)$. Our construction has the property that $(\sigma(D \mid m = (1, ..., 1))$ and $(\sigma(D \mid m = (0, ..., 0))$ are both the same as the distribution over D conditional on $\mathcal{B}^*$. Consequently, the strategy satisfies the independence property $D \perp \theta$ (on top of the property $D \perp \theta \mid m$ that was established by Lemma 1). Thus, even if the receiver attempted to draw inferences from D, he would be unable to learn anything about θ from the realization of D itself.

As to the question of how large dictionaries should be (discussed in the context of Example 2), note that the sender's optimal strategy makes use of dictionaries that interprets exactly *half* of the message components.

Let us examine the receiver's reaction to various realized reports under the sender's strategy. When he confronts the message $(0, ..., 0)$, each of the dictionaries that accompany it interprets some "bad news", and the receiver learns that $\theta = N$ for sure. In contrast, every other realization of (m, D) satisfies $m_k = 1$ for all $k \in D$. The receiver thus learns that the probability of m_D conditional on $\theta = Y$ is one, while the probability of m_D conditional on $\theta = N$ is $\min\{1/S, \pi/(1-\pi)\}$. The receiver's subjective likelihood ratio of (m, D) is

$$\rho_\sigma(m, D) = \frac{1}{\min\{\frac{1}{S}, \frac{\pi}{1-\pi}\}}$$

which is, by definition, weakly above $(1-\pi)/\pi$ and therefore persuasive.

A receiver with rational expectations would realize that the "mixed" messages in $\mathcal{B}^*$ only occur in state N. However, our receiver can only draw inferences from message components that the sender interprets for him. Since the sender only interprets persuasive patterns, he manages to convey a false sense that the mixed message is actually good news. Moreover, as K gets large, each $(m, D) \in \mathcal{B}^*$ identifies a distinct pattern that becomes increasingly rare in state N while occurring with probability one in state Y. Therefore, even when π is quite small and even if $\mathcal{B}^*$ is played with high probability in state N, the receiver will be persuaded by the reports in $\mathcal{B}^*$.

When $\pi \geq 1/(1+S)$, the sender can attain full persuasion. This means that the sender's strategy is *time-consistent*: Since the receiver plays $a = y$ after every report, the sender would not want to deviate from any realized report even if he could. In other words, the assumption that the sender has commitment power is not required in this range of parameters.

Theorem 1 assumes an unrestricted domain of feasible dictionaries. The proof of Theorem 1 makes the result easily extendible to restricted domains.

Remark 2 *Let $\mathcal{D}$ be the set of feasible dictionaries. Let $\mathcal{D}^* \subseteq \mathcal{D}$ be an anti-chain, such that every $\mathcal{D}' \subseteq \mathcal{D}$ with $|\mathcal{D}'| > |\mathcal{D}^*|$ is not an anti-chain. Then, the maximal probability of persuasion is* $\min\{1, \pi(1 + |\mathcal{D}^*|)\}$.

In particular, when the feasible set of dictionaries is the set of all *singletons*, the maximal probability of persuasion is $\max\{1, \pi(1 + K)\}$. This suggests that if the sender were free to determine the dimensionality of the message space, he could trivially attain full persuasion with singleton dictionaries. However, K should be interpreted as an exogenous constraint: there is a *limited* set of variables about which statistical data is available. For instance, if message components correspond to non-verbal aspects of the sender's communication, only few of those aspects are typically documented (it is unlikely to have data about the sender's pupil dilation, blood pressure or EEG measurements). Similarly, if the sender is a political party and message components correspond to different party members, only the messages of a few senior members are likely to be documented.

4 Suspicion of Selective Interpretations

In our discussion of Theorem 1, we raised the concern that the receiver may try to infer the state from the dictionary the sender provides. The sender strategy we presented in the theorem's statement addressed this concern, in the sense that it satisfied the independence property $D \perp \theta$. However, one may argue that even this feature would not quell the receiver's suspicion regarding the *selectiveness* of the provided dictionary - i.e., some message components are not interpreted. The receiver may view the mere neglect of message components as a signal that the state is N (even though the state-contingent distribution over dictionaries offers no basis for this suspicion).

While intuitive, this argument is actually unconventional. The receiver draws a correct Bayesian inference from the message components for which he gets data. In the absence of additional data on how dictionaries and messages are jointly distributed, there is nothing to guide the receiver on how to modify this Bayesian posterior. Any assertion that he should ignore his available data and conclude that the state must be N simply because he was given selective data by a strategic sender is merely an *additional assumption*. By the same token, one could argue that in the partially informative "interval equilibria" in Crawford and Sobel (1982), the receiver should ignore his statistical knowledge of the sender's behavior and trust *nothing* the sender says simply because he is known to lie or withhold information.[4]

This methodological discussion notwithstanding, we now address the possibility that receivers may be suspicious of selective interpretations by proposing two notions of robustness to this suspicion. In both cases, we show that full persuasion is attainable for a large range of parameters π, K, albeit smaller than in Theorem 1.

[4] If we interpret the sender's strategy as recommending an action or communicating the interval to which the state belongs, this is a case of witholding information. If we interpret his strategy as some mixture over states that belong to the interval, then his message misrepresents the state with probability one.

4.1 Benevolent Selectiveness

Even if the sender's interests were fully aligned with the receiver's, it would be reasonable for him to refrain from interpreting *all* message components and provide a selective dictionary. To see why, let $K = 2$ and suppose that the message strategy is as follows: $m = (1, 1)$ with certainty in state Y, whereas $m = (0, 0)$ and $m = (1, 1)$ with equal probability in state N. Because m_1 and m_2 are fully correlated, the small dictionary $\{1\}$ induces the same receiver beliefs as the full dictionary $\{1, 2\}$. If the smaller dictionary is less costly to provide, a benevolent sender would use it (recall that D must be non-empty). In this case, the receiver would not be suspicious of the sender simply for providing a small dictionary.

To capture this idea, we modify our model by introducing an intrinsic preference for smaller dictionaries. Specifically, we assume that the sender has lexicographic preferences. His primary criterion is to maximize the probability that the receiver plays y. However, if he can induce the same receiver behavior with two alternative dictionaries D and D' such that $|D'| < |D|$, he prefers D' to D. In addition, we impose a refinement of the set of permissible sender strategies, which is based on a hypothetical *benevolent* sender. Such a sender has lexicographic preferences, too: His primary criterion is to maximize the receiver's payoff; his secondary criterion is to minimize $|D|$. Refer to this *hypothetical* sender as type H; whereas the *actual* sender will be referred to as type A.

Definition 2 *The strategy $(\sigma(m, D \mid \theta))$ is **robust** if it satisfies the following properties:*
(i) $D \perp \theta$ and $D \perp \theta \mid m$.
(ii) Given $(\sigma(m \mid \theta))$, the interpretation strategy $(\sigma(D \mid m))$ prescribes, for each m, lexicographically optimal dictionaries for a type-A sender.
(iii) Given $(\sigma(m \mid \theta))$, there is an interpretation strategy $(\sigma'(D \mid m))$ that prescribes, for each m, lexicographically optimal dictionaries for a type-H sender, such that $\sigma'(D) \equiv \sigma(D)$.

Condition (i) imposes the independence requirements we have already encountered in Section 3. Condition (*ii*) was redundant in Section 3 because we focused on optimal sender strategies anyhow. Here, it also means that the sender always uses the smallest dictionary that attains a given outcome.

As to condition (*iii*), our motivation is the following. Throughout the paper, we have assumed that the receiver lacks any data about the distribution of D. However, imagine now that the receiver has access to an independent dataset that enables him to learn the marginal distribution of dictionaries. (By condition (*i*), this is the same as learning the distribution of D at each state.) He can therefore see that the use of selective dictionaries is not a fluke, but an event that occurs with positive frequency. Condition (*iii*) requires further that if the dictionaries were chosen by a benevolent sender of type H, their marginal distribution could be the same. From this point of view, the receiver is less likely to be suspicious of selective interpretations, because he can reconcile their observed statistical pattern with the existence of a benevolent interpreter having a lexicographically secondary preference for small dictionaries.

In what follows, we conveniently assume that the receiver always breaks ties in favor of a type-A sender.

Proposition 2 *Full persuasion is attainable with a robust strategy if and only if* $\pi \geq 1/(1+K)$.

Thus, requiring the sender's strategy to be robust in the sense of Definition 2 restricts his ability to attain full persuasion, because it effectively eliminates the use of non-singleton dictionaries. Example 2 in Section 3 illustrates a robust strategy that achieves full persuasion for $K = 3$.

4.2 Full-Coverage Dictionaries

In this subsection we use a different line of attack to address the selective-interpretation problem. Here, we assume that the sender is obliged to present statistical data about *all* message components. However, he is allowed to present the data in two separate chunks. As before, a dictionary is represented by a non-empty subset $D \subseteq \{1, ..., K\}$.

Yet this now means that the sender provides *two* datasets, formalized as two collections of conditional probabilities: $(\Pr(m_D \mid \theta))$ as well as the $(\Pr(m_{D^c} \mid \theta))$, where $D^c = \{1, ..., K\} \backslash D$. We refer to this form of data provision as *full-coverage dictionaries.*

How does the receiver extrapolate a belief from the two datasets? We make the mild assumption that his subjective belief $\widetilde{\Pr}(m, D \mid \theta)$ satisfies

$$\Pr(m_D \mid \theta) \cdot \Pr(m_{D^c} \mid \theta) \leq \widetilde{\Pr}(m, D \mid \theta) \leq \max\{\Pr(m_D \mid \theta), \Pr(m_{D^c} \mid \theta)\} \tag{4}$$

The upper bound given by the R.H.S reflects an assumption that m_{D^c} is uninformative of θ given m_D, or vice versa - i.e., the two parts of m are perfectly correlated given the state. The lower bound given by the L.H.S reflects an assumption that these two parts are independent conditional on the state.

Proposition 3 *Let $K > 2$. Then, the sender can attain full persuasion with full-coverage dictionaries whenever*

$$\pi \geq \frac{4}{4+S}$$

Thus, although the sender is forced to provide data about *all* message components, his ability to present the data in two "installments" enables him to attain full persuasion for a large range of parameters. Moreover, the strategy we construct in the proof satisfies the familiar independence properties $D \perp \theta$ and $D \perp \theta \mid m$. Finally, the result relies on the relatively weak condition (4) on how the receiver extrapolates a belief from the two separate datasets he receives. Note that Proposition 3 only provides a sufficient condition for full persuasion. Finding a necessary condition is an open problem.

To illustrate the basic idea of the construction, let $K = 4$. In state Y, there is perfect correlation among all message components. The objective correlation is weaker in state N. Specifically, only the messages $(1, 1, 1, 1)$ and $(0, 0, 0, 0)$ are played in Y, whereas all messages containing exactly two 1's are played in N. Thus, patterns like $(*, 1, 1, *)$ or $(0, *, *, 0)$ are considerably more likely in Y than in N. By accompanying the message $(0, 1, 1, 0)$ with two datasets that separately highlight these two patterns, the sender can manipulate the receiver's likelihood ratio.

This subsection also illustrates that the form a dictionary can affect the sender's ability to persuade the receiver. This reinforces a point we made in Section 2: Our concept of "selective interpretation" is richer than what the "selective message redaction" metaphor might suggest.

5 Richer Dictionaries

In this section we follow up on the final paragraph of the previous section. So far, we have assumed that dictionaries provide data about the joint distribution of a collection of message components conditional on θ. However, statistical data can involve other combinations of marginal and conditional distributions, with implications for the sender's ability to persuade the receiver.

Example 3
Let $K = 2$. Let p denote the joint distribution over (θ, m) that is induced by the prior over θ and the sender's strategy. There are three feasible dictionaries: D_1 gives access to the conditional distribution $(p(m_1 \mid \theta))$; D_2 gives access to the conditional distribution $(p(m_2 \mid \theta))$; and D_3 gives access to the marginal distribution $(p(m_1))$ *as well as* the conditional distribution $(p(m_2 \mid \theta, m_1))$. It does not contain data about how m_1 varies with θ.

The dictionaries D_1 and D_2 are familiar from Section 2; we apply the same belief-formation rule (1) for the receiver as in Section 2. However, D_3 is different because it provides *two* datasets. We assume that the receiver extrapolates a belief using the *maximum entropy principle* - i.e., his belief over (θ, m_1, m_2) maximizes (Shannon) entropy subject to the constraint that it is consistent with the marginal and conditional distributions he has learned. This principle has a rich tradition in AI (dating back to Jaynes (1957)). Spiegler (2020) has recently applied it in a similar context of games with players who extrapolate a belief from partial data. In the model of Section 4.2, the principle induces the L.H.S of (4). In the present context, the receiver's subjective distribution over messages conditional on the state, given D_3, is $\widetilde{\Pr}(m_1, m_2 \mid \theta) = p(m_1)p(m_2 \mid \theta, m_1)$.

Consider the following sender strategy:

State Y

m	D	$\Pr(m, D \mid Y)$
$(1,1)$	D_3	ε
$(0,0)$	D_2	$1-\varepsilon$

State N

m	D	$\Pr(m, D \mid Y)$
$(1,1)$	D_3	α
$(1,0)$	D_2	β
$(0,1)$	D_1	$1-\alpha-\beta$

We now show that for every $\pi > \frac{1}{10}\left(5-\sqrt{5}\right)$, there exist $\alpha, \beta, \varepsilon \in (0,1)$ such that the sender attains full persuasion with the above strategy.

Let us calculate the receiver's likelihood ratio for each report. Consider the report $((1,1), D_3)$. Our definition of the receiver's posterior belief given the dictionary D_3 implies the following likelihood ratio:

$$\frac{p(m_1 = 1)p(m_2 = 1 \mid \theta = Y, m_1 = 1)}{p(m_1 = 1)p(m_2 = 1 \mid \theta = N, m_1 = 1)} = \frac{1}{\frac{\alpha}{\alpha+\beta}} = \frac{\alpha+\beta}{\alpha}$$

Next, consider the reports $((0,0), D_2)$ and $((1,0), D_2)$. Since D_2 only interprets m_2, both reports induce the same subjective likelihood ratio:

$$\frac{p(m_2 = 0 \mid \theta = Y)}{p(m_2 = 0 \mid \theta = N)} = \frac{1-\varepsilon}{\beta}$$

Finally, consider the report $(0, D_1)$. Since D_1 only interprets m_1, this report induces the subjective likelihood ratio

$$\frac{p(m_1 = 0 \mid \theta = Y)}{p(m_1 = 0 \mid \theta = N)} = \frac{1-\varepsilon}{1-\alpha-\beta}$$

In order to attain full persuasion, the three likelihood ratios must all be weakly greater than $(1-\pi)/\pi$. A straightforward calculation establishes that whenever

$\pi > \frac{1}{10}\left(5 - \sqrt{5}\right)$, we can find $\alpha, \beta, \varepsilon$ that will satisfy these three inequalities. In particular, ε will be arbitrarily small. □

Compare this finding with the result of Section 3. Given our original specification of dictionaries, the sender can attain full persuasion if and only if $\pi \geq \frac{1}{3}$. This is *higher* than the threshold we obtained in Example 3. The general problem of optimal persuasion under the broader definition of dictionaries as collections of marginal and conditional distributions remains open.

6 An Adversarial Sender

In this section we revisit the basic model of Section 2 and modify the sender's preferences, such that the sender-receiver interaction becomes a zero-sum game: In state Y (N), the sender's payoff is 1 if the receiver plays n (y) and -1 if he plays y (n). Rescale the receiver's payoff function to be minus the sender's payoff. In what follows, we assume that the receiver always breaks ties in the sender's favor.

Consider the rational-expectations benchmark in this case. On one hand, the receiver can guarantee an expected payoff of at least $\pi \cdot (-1) + (1 - \pi) \cdot 1 = 1 - 2\pi > 0$ by always playing n. On the other hand, the sender can force this expected payoff on the receiver by sending the same report in all states. Therefore, by the Minimax Theorem, the sender's equilibrium payoff in the rational-expectations benchmark is exactly $2\pi - 1 < 0$. In contrast, the following result establishes that in our model, the sender can attain the maximal possible payoff of 1 under the same condition as in Theorem 1, whenever $K \geq 3$.

Proposition 4 *Let $K \geq 3$. Then, whenever $\pi \geq 1/(1 + S)$, there is a strategy for the sender that induces a payoff of* 1 *with certainty.*

Proof. Construct the following strategy. Let $m_k \in \{0, 1\}$ for every k. In state Y, the sender plays $m^* = (1, 1, ..., 1)$ with probability one and accompanies this message with the dictionary $D = \{k\}$ for some arbitrary k. In state N, the sender assigns probability $(1 - \gamma)/S$ to every (m, D) satisfying $m_k = 1$ for exactly $\lfloor K/2 \rfloor$

components k and $D = \{k \mid m_k = 1\}$, where γ is selected to be the unique solution of the equation

$$\frac{1}{\gamma + \frac{1}{S}(1-\gamma)} = \frac{1-\pi}{\pi}$$

The sender assigns the remaining probability γ to the message m^* and accompanies it with an arbitrary dictionary of size $\lfloor K/2 \rfloor$. This is a feasible strategy whenever $\gamma \in [0,1]$ or equivalently $\pi \in \left[1/(1+S), \frac{1}{2}\right]$.

By construction, $\rho(m,D) = (1-\pi)/\pi$ for every (m,D) that is played in state N, whereas

$$\rho(m^*, \{k\}) = \frac{1}{\gamma + \frac{1}{2}(1-\gamma)} < \frac{1-\pi}{\pi}$$

As a result, the receiver plays y in state N and n in state Y, generating a payoff of 1 for the sender. ■

Thus, strategic interpretation can attain the sender's first-best even under maximal conflict of interests with the receiver. As in Section 3, this means that the commitment assumption is unnecessary.

However, the strategy we employed in the proof of this result violates two independence properties that we emphasized in Section 3: $D \perp \theta$ and $D \perp \theta \mid m$. Let us now see how to fix this limitation when $K \geq 3$ and $\pi \geq 1/K$. As before, $m_k \in \{0,1\}$ for every k. Let e_k denote the message m for which $m_k = 1$ and $m_l = 0$ for all $l \neq k$. For every m, let $-m$ denote the message m' for which $m'_k = 1 - m_k$ for every k. Now consider the following sender strategy. In state Y, he randomizes uniformly over all (m,D) such that $m = -e_k$ and $D = \{k\}$ for some $k = 1, ..., K$. In state N, he randomizes uniformly over all (m,D) for which $m = e_k$ and $D = \{k\}$ for some $k = 1, ...K$. It is easy to verify that $\rho(m,D) \geq (1-\pi)/\pi$ for every (m,D) that is played in N, while $\rho(m,D) \leq (1-\pi)/\pi$ for every (m,D) that is played in Y, as long as $\pi \geq 1/K$.

The following result expands the set of parameters for which the sender's first-best is attainable by a strategy that satisfies the two desiderata, making use of a more elaborate strategy.

Proposition 5 *Let $K = 2L$ for some integer $L > 1$. Then, there is a strategy that satisfies $D \perp \theta \mid m$ and $D \perp \theta$ and attains the sender's first-best whenever*

$$\pi \geq \frac{1}{1 + \binom{L}{\lfloor L/2 \rfloor}}$$

This result provides a sufficient condition for attaining the sender's first-best with a strategy that satisfies the two desiderata. The following table illustrates the strategy for $K = 4$ (the strategy induces the sender-optimal action in each state, as long as $\pi \geq \frac{1}{3}$):

State Y			State N		
m	D	$\Pr(m, D \mid Y)$	m	D	$\Pr(m, D \mid Y)$
0011	$\{1\}$	0.25	1000	$\{1\}$	0.25
0011	$\{2\}$	0.25	0100	$\{2\}$	0.25
1100	$\{3\}$	0.25	0010	$\{3\}$	0.25
1100	$\{4\}$	0.25	0001	$\{4\}$	0.25

Finding a tight necessary condition remains an open problem.

7 Related Literature

Our paper joins a small literature on strategic communication that departs from the standard paradigm of rational expectations under a common prior. Levy et al. (2018) study a sender-receiver model in which the receiver exhibits "correlation neglect". Specifically, the sender submits multiple simultaneous signals and the receiver erroneously treats them as being conditionally independent. This belief distortion is related to the model of Section 4.2. In that variant on our basic model, the receiver does not learn the state-contingent correlation between m_D and m_{D^c}. We allowed the receiver to hold a variety of beliefs regarding this correlation, including the possibility that they are conditionally independent, as in Levy et al. (2018). The reason that unlike Levy et al. (2018), the sender in our model can attain full persuasion is that he can tailor the data he gives the receiver to the submitted message.

Patil and Salant (2020) consider a receiver (a statistician) who estimates a parameter based on a random sample whose size is strategically determined by an informed sender. As in our model, the receiver has boundedly rational expectations in the sense that he makes no inferences from the sample size he gets. Schwartzstein and Sunderam (2019) examine a persuasion game in which both parties observe a signal that is drawn from a state-dependent distribution. The receiver's non-rational expectations are captured by the assumption that the sender knows the signal distribution, while the receiver believes in whatever signal distribution the sender reports. Galperti (2019) analyses a model of persuasion with non-common priors, where the sender can influence the receiver's prior belief. In particular, when the receiver observes a message that has zero probability according to his prior, he abandons it in favor of a new belief. We, on the other hand, maintain the common prior assumption but allow the sender to strategically determine the receiver's understanding of the equilibrium distribution.[5]

Our basic model of dictionaries and how the receiver reacts to them is closely related to the concept of analogy-based expectations equilibrium (ABEE) due to Jehiel (2005). According to this concept, players form coarse beliefs that are measurable with respect to an "analogy partition" of the possible states of the world. Our basic notion of a dictionary D as a subset of components of multi-dimensional messages corresponds to an analogy partition. A cell in the partition consists of all messages m with the same m_D. This version of the model can thus be viewed as an extensive game in which the sender chooses the message as well as the receiver's analogy partition (from a restricted domain of feasible partitions), and the solution concept is ABEE. (However, the variants of Sections 4.2 and 5 *cannot* be embedded in the ABEE framework.) This description raises a natural question: How well can the sender perform under an *unrestricted* domain of feasible analogy partitions? For the sake of brevity, we do not analyze this question here but in a separate note (Eliaz

[5] Independently of our paper, Salcedo (2019) considers a persuasion game with one sender who sends private messages to multiple *rational* receivers. The sender wishes to persuade at least m receivers in order to attain his objective. When $m = 1$, the sender's problem is essentially the same as the sender's problem in our model when he is restricted to singleton dictionaries.

et al. (2019)).

Jehiel and Koessler (2008) modify the Crawford-Sobel model by assuming that the receiver bundles states into analogy classes according to an interval analogy partition. They show that certain analogy partitions give rise to ABEE with partial information transmission, even when the unique equilibrium under rational expectations is the babbling equilibrium. Hagenbach and Koessler (2019) analyze cheap-talk games where the sender aggregates the receiver's equilibrium strategy into analogy classes. In a similar vein, Mullainathan et al. (2008) study a cheap-talk game where the receiver uses a coarse analogy partition. In contrast to our model, the partitions in these papers are exogenous. Endogenous partitions arise in Jehiel (2011), where auction designer controls bidders' learning feedback regarding the distribution of past bids.

Glazer and Rubinstein (2012, 2014) study persuasion when the *sender* is boundedly rational in the sense of having limited ability to misrepresent the state. They show that a rational receiver can construct intricate disclosure mechanisms that take advantage of this element of the sender's bounded rationality. Blume and Board (2013) and Giovannoni and Xiong (2019) study cheap talk when the receiver has uncertain ability to distinguish between distinct messages. In contrast to our framework, receivers in these papers have rational expectations and the sender is unable to influence their interpretative abilities.

Finally, Spiegler (2020) introduces a general framework for static games, in which the description of players' types includes "archival access", defined as selective data about correlations among the variables that constitute the state of the world. Dictionaries in our model are a form of archival access. Indeed, our model is an example of how to extend the formalism of Spiegler (2020) to sequential games. Our approach to modeling the receiver's partial understanding of the sender's strategy is also related to Glazer and Rubinstein (2019), where a "problem solver" has partial understanding of the equilibrium: He observes a summary statistic of the other players' strategies, and then best-replies to a uniform belief over all the strategy profiles that are consistent with this statistic.

8 Conclusion

Conventional models of strategic communication focus on the role of selective transmission of information. And yet, real-life communication also involves strategic *interpretation* of information. This paper formalized this aspect as selective provision of *statistical data* regarding the mapping from states to messages, under the assumption that this data is the sole basis for the receiver's inferences. In a pure persuasion model, we showed that strategic interpretation significantly enhances the sender's ability to persuade the receiver - to the point that *full* persuasion is sometimes possible, in sharp contrast to the standard rational-expectations benchmark.

From a broader perspective, the modeling innovation in this paper is the idea that one player can influence another player's understanding of equilibrium regularities, by affecting the statistical data regarding the equilibrium distribution that the latter player has at his disposal (his "archival access", to use the terminology of Spiegler (2020)) - just as in a standard extensive-form game, one player's information set can be determined by prior moves of other players. Exploring this idea outside the context of strategic communication is an interesting problem for future research.

References

Andreas Blume and Oliver Board. Language barriers. *Econometrica*, 81(2):781–812, 2013.

Vincent P Crawford and Joel Sobel. Strategic information transmission. *Econometrica*, pages 1431–1451, 1982.

Kfir Eliaz, Ran Spiegler, and Heidi C. Thysen. Strategic interpretations. 2018.

Kfir Eliaz, Ran Spiegler, and Heidi C. Thysen. On persuasion with endogenous misspecified beliefs. 2019.

Simone Galperti. Persuasion: The art of changing worldviews. *American Economic Review*, 109(3):996–1031, 2019.

Francesco Giovannoni and Siyang Xiong. Communication under language barriers. *Journal of Economic Theory*, 180:274–303, 2019.

Jacob Glazer and Ariel Rubinstein. On optimal rules of persuasion. *Econometrica*, 72(6):1715–1736, 2004.

Jacob Glazer and Ariel Rubinstein. A study in the pragmatics of persuasion: A game theoretical approach. *Theoretical Economics*, 1:395–410, 2006.

Jacob Glazer and Ariel Rubinstein. A model of persuasion with boundedly rational agents. *Journal of Political Economy*, 120(6):1057–1082, 2012.

Jacob Glazer and Ariel Rubinstein. Complex questionnaires. *Econometrica*, 82(4): 1529–1541, 2014.

Jacob Glazer and Ariel Rubinstein. Coordinating with a ”problem solver”. *Management Science*, 65:2813–2819, 2019.

Jeanne Hagenbach and Frédéric Koessler. Cheap talk with coarse understanding. mimeo, 2019.

Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106 (4):620, 1957.

Philippe Jehiel. Analogy-based expectation equilibrium. *Journal of Economic theory*, 123(2):81–104, 2005.

Philippe Jehiel. Manipulative auction design. *Theoretical economics*, 6(2):185–217, 2011.

Philippe Jehiel and Frédéric Koessler. Revisiting games of incomplete information with analogy-based expectations. *Games and Economic Behavior*, 62(2):533–557, 2008.

Ginger Zhe Jin, Michael Luca, and Daniel Martin. Is no news (perceived as) bad news? An Experimental Investigation of Information Disclosure, 2019.

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

Gilat Levy, Inés Moreno de Barreda, and Ronny Razin. Persuasion with correlation neglect: Media power via correlation of news content. 2018.

Sendhil Mullainathan, Joshua Schwartzstein, and Andrei Shleifer. Coarse thinking and persuasion. *The Quarterly journal of economics*, 123(2):577–619, 2008.

Sanket Patil and Yuval Salant. Persuading statisticians. 2020.

Bruno Salcedo. Persuading part of an audience. 2019.

Joshua Schwartzstein and Adi Sunderam. Using models to persuade. December 2019.

Ran Spiegler. Modeling players with random "data access". July 2020.

Appendix: Proofs

Proposition 1

Let σ be an optimal sender strategy. We now change it into a new strategy that satisfies the property in the statement of the proposition and does not lower the probability of persuasion. We proceed in two stages.

Stage 1. Construct a partition $\{T_1, ..., T_L\}$ of $\mathcal{B}_\sigma$ as follows. For every $l = 1, 2, ...$, select an arbitrary report $(m^l, D^l) \in \mathcal{B}_\sigma - \cup_{h<l} T_h$, and define

$$T_l = \{(m, D) \in \mathcal{B}_\sigma - \cup_{h<l} T_h \mid m_{D^l} = m^l_{D^l}\}$$

Modify σ as follows. For each $l = 1, ..., L$ and any $(m, D) \in T_l$ with $D \neq D^l$, shift the probability of (m, D), conditional on $\theta = N$, to the report (m, D^l). By the definition of $\mathcal{B}_\sigma$, both (m, D) and (m^l, D^l) persuade the receiver. Perform the following additional modification. By the definition of $\mathcal{B}_\sigma$, there must be a message m that justifies (m^l, D^l). That is, $m_{D^l} = m^l_{D^l}$, and there is a dictionary D such that (m, D) is played with positive probability in Y. If the receiver was persuaded by (m, D) in the original strategy, then shift the probability of every such (m, D) conditional on Y to (m, D^l). By construction, $m_{D^l} = m^l_{D^l}$. Therefore, (m, D^l) persuades the receiver. And since the deviation does not affect the distribution over messages conditional on any state, it does not change the receiver's response to any other realized report.

Stage 2. Start this stage by shifting the probability of any $(m, D^L) \in T_L$ conditional on $\theta = N$ to some report in T_L, denoted $(\tilde{m}^L, D^L)$. This effectively transforms T_L into a singleton $\{(\tilde{m}^L, D^L)\}$. By the construction of the first phase, every $(m, D^L) \in T_L$ satisfies $m_{D^L} = \tilde{m}^L_{D^L}$. Therefore, the deviation does not change the receiver's subjective likelihood ratio of $(\tilde{m}^L, D^L)$, such that he continues to be persuaded by this report. Moreover, by the construction of the first stage, for every $l < L$ and every $(m, D^l) \in T_l$, $m_{D^l} \neq \tilde{m}^L_{D^l}$. Therefore, the deviation does not affect the receiver's subjective likelihood ratio of $(m, D^l) \in T_l$ for all $l < L$.

Now suppose that for some $l < L$, we have transformed the cells $T_{l+1}, ..., T_L$ into singletons $\{(\tilde{m}^{l+1}, D^{l+1})\}, ..., \{(\tilde{m}^L, D^L)\}$ in such a manner. Suppose that there is some $(m, D^l) \in T_l$ such that $m_{D^h} \neq \tilde{m}^h_{D^h}$ for every $h > l$. Rename this report $(\tilde{m}^l, D^l)$, and shift the probability of any (m, D^l) conditional on N to $(\tilde{m}^l, D^l)$. Alternatively, suppose that for every $(m, D^l) \in T_l$ there is some $h > l$ such that $m_{D^h} = \tilde{m}^h_{D^h}$. For any such (m, D^l), shift its probability conditional on N to one of the reports $(\tilde{m}^h, D^h)$ satisfying $\tilde{m}^h_{D^h} = m_{D^h}$. By the same logic as in the previous paragraph, the deviation in these two alternative cases does not affect the receiver's subjective likelihood ratio of any report.

At the end of the second stage, $\mathcal{B}_\sigma$ has been effectively transformed into the set $\{(\tilde{m}^1, D^1)\}, ..., \{(\tilde{m}^L, D^L)\}$, which by construction satisfies the property in the lemma's statement.

In the next two corollaries, we restrict attention to sender strategies σ that satisfy Proposition 1.

Corollary 1

Assume, by contradiction, that there exist $(m, D), (m', D') \in \mathcal{B}_\sigma$ that are justified by a message m^* and $D \subseteq D'$. This means that $m^*_D = m_D$ and $m^*_{D'} = m'_{D'}$. Therefore, $m_{D \cap D'} = m^*_{D \cap D'} = m'_{D \cap D'}$. But $D \cap D' = D$, which implies that $m_D = m'_D$, in contradiction to Proposition 1.

Corollary 2

By Corollary 1, if m^* justifies two reports (m, D) and (m', D'), then D and D' do not contain one another. It follows that the set of all dictionaries that are part of reports justified by m^* constitutes an anti-chain - i.e. a collection of subsets of $\{1, ..., K\}$ that do not contain one another. By Sperner's Theorem, the maximal size of such a collection is S.

Theorem 1

To derive an upper bound on the probability of persuasion, we restrict attention to sender strategies σ that satisfy Proposition 1. We begin with a basic observation that simplifies notation and the construction of the sender's strategy that maximizes the probability of persuasion in the N event. Fix a sender' strategy.

Observation 1 *There is no loss of generality in restricting attention to strategies with the following property: If the reports $(m, D) \in \mathcal{B}_\sigma$ and $(m', D') \notin \mathcal{B}_\sigma$ are both realized with positive probability in the N state under σ, then $m'_D \neq m_D$.*

Proof. Assume the contrary - i.e. $m'_D = m_D$. Suppose the sender deviates to a strategy that replaces (m', D') with (m', D) in the N state, but otherwise coincides with σ. By definition of $\mathcal{B}_\sigma$, (m', D') does not persuade the receiver prior to the deviation. And since the deviation does not affect the distribution of messages conditional on any state, it does not change the response of the receiver to any report $(m'', D'') \neq (m', D')$. Therefore, the deviation weakly raises the probability of persuasion. ■

Henceforth, we will restrict attention to strategies that satisfy Observation 1. In addition, whenever we refer to a generic report in the N state, we mean a report in $\mathcal{B}_\sigma$.

Lemma 2 *Without loss of generality, $\rho_\sigma(m, D)$ is the same for all $(m, D) \in \mathcal{B}_\sigma$.*

Proof. Let $(\underline{m}, \underline{D})$ and $(\bar{m}, \bar{D})$ be two reports in $\mathcal{B}_\sigma$ such that $\rho_\sigma(\underline{m}, \underline{D}) \leq \rho_\sigma(m, D) \leq \rho_\sigma(\bar{m}, \bar{D})$ for each $(m, D) \in \mathcal{B}_\sigma$. Assume that $\rho_\sigma(\underline{m}, \underline{D}) < \rho_\sigma(\bar{m}, \bar{D})$. Suppose that the sender deviates from σ to a strategy $\hat{\sigma}$ that shifts a weight of $\varepsilon > 0$ from $(\underline{m}, \underline{D})$ to $(\bar{m}, \bar{D})$ in state N. By Proposition 1, $\bar{m}_{\underline{D}} \neq \underline{m}_{\underline{D}}$ and $\underline{m}_{\bar{D}} \neq \bar{m}_{\bar{D}}$. Therefore,

$$\rho_{\hat{\sigma}}(\underline{m}, \underline{D}) = \frac{\sum_{m|m_{\underline{D}}=\underline{m}_{\underline{D}}} \sigma(m \mid \theta = Y)}{\sum_{m|m_{\underline{D}}=\underline{m}_{\underline{D}}} \sigma(m \mid \theta = N) - \varepsilon} > \rho_\sigma(\underline{m}, \underline{D}) \geq \frac{1-\pi}{\pi} \tag{5}$$

$$\rho_{\hat{\sigma}}(\bar{m}, \bar{D}) = \frac{\sum_{m|m_{\bar{D}}=\bar{m}_{\bar{D}}} \sigma(m \mid \theta = Y)}{\sum_{m|m_{\bar{D}}=\bar{m}_{\bar{D}}} \sigma(m \mid \theta = N) + \varepsilon} < \rho_\sigma(\bar{m}, \bar{D})$$

By our initial assumption, $\rho_{\hat{\sigma}}(\underline{m}, \underline{D}) < \rho_{\hat{\sigma}}(\bar{m}, \bar{D})$ for sufficiently small ε. By (5), this implies that $\rho_{\hat{\sigma}}(\bar{m}, \bar{D}) > \frac{1-\pi}{\pi}$. By Proposition 1 $\rho_{\hat{\sigma}}(m, D) = \rho_\sigma(m, D)$ for every $(m, D) \in \mathcal{B}_\sigma - \{(\underline{m}, \underline{D}), (\bar{m}, \bar{D})\}$. Since the deviation does not involve reports outside $\mathcal{B}_\sigma$, it cannot alter the probability of persuading the receiver for messages outside of $\mathcal{B}_\sigma$. It follows that the deviation does not alter the probability of persuasion.

Therefore, we can assume without loss of generality that $\rho_\sigma(m, D)$ is the same for all $(m, D) \in \mathcal{B}_\sigma$. ■

The remainder of the proof computes an upper bound on the probability of persuasion. Let σ be a sender strategy. Let $\mathcal{M}_Y = \{m \mid \sigma(m \mid \theta = Y) > 0\}$. Denote $I = |\mathcal{M}_Y|$. Let $\mathcal{C} = \{C_1, \cdots, C_L\}$ be a partition of $\mathcal{B}_\sigma$, where each cell C_l is defined by the (distinct) subset of messages $J(l) \subseteq \mathcal{M}_Y$ that justify every report in the cell. Therefore, $L \leq 2^I - 1$. For the final piece of notation we let $g(l) =\mid C_l \mid$ and $\beta(l) = \sum_{(m,D)\in C_l} \sigma(m, D \mid \theta = N)$.

Consider some $(m, D) \in C_l \subseteq \mathcal{B}_\sigma$ and a message $m' \in J(l)$. Since m' justifies (m, D), $m'_D = m_D$. By Proposition 1, there cannot be a dictionary D' such that $(m', D') \in \mathcal{B}_\sigma$. It follows that for any $l = 1, ..., L$, the receiver's subjective likelihood ratio of a report $(m, D) \in C_l \subseteq \mathcal{B}_\sigma$ is

$$\frac{\sum_{m' \in J(l)} \sigma(m' \mid \theta = Y)}{\sigma(m, D \mid \theta = N)} \geq \frac{1-\pi}{\pi}. \tag{6}$$

From lemma 2 we have $\rho(m, D) = \rho(m', D')$ for every $(m, D), (m'D') \in \mathcal{B}_\sigma$. So in particular for every $(m, D), (m'D') \in C_l$ we have $\sigma(m, D) = \sigma(m', D') = \frac{\beta(l)}{g(l)}$. We can therefore rewrite inequality 6 as:

$$\frac{\sum_{m' \in J(l)} \sigma(m' \mid \theta = Y)}{\frac{\beta(l)}{g(l)}} \geq \frac{1-\pi}{\pi}, \tag{7}$$

Solving for $\beta(l)$ in (7) and summing over l give us

$$\begin{aligned} \sum_{l=1}^{L} \beta(l) &\leq \sum_{l=1}^{L} g(l) \sum_{m' \in J(l)} \left[\frac{\pi}{1-\pi} \sigma(m' \mid \theta = Y) \right] \\ &= \sum_{m' \in M^*} \left[\frac{\pi}{1-\pi} \sigma(m' \mid \theta = Y) \right] \sum_{l \in J^{-1}(m')} g(l) \end{aligned}$$

where the second equality follows from changing the order of summation. By definition, $\sum_{l \in J^{-1}(m')} g(l)$ is the number of reports that are justified by m'. By Corollary 2, this number is at most S. Therefore,

$$\begin{aligned} \sum_{l=1}^{L} \beta(l) &\leq \sum_{m' \in M^*} \left[\frac{\pi}{1-\pi} \sigma(m' \mid \theta = Y) \right] S \\ &= \frac{\pi}{1-\pi} S \end{aligned} \tag{8}$$

where the final equality follows since $\sum_{m' \in M^*} \sigma(m' \mid \theta = Y) = 1$. Since the receiver can at most be persuaded with probability one, the upper bound on the probability of persuasion in the N state is

$$\min \left\{ \frac{\pi}{1-\pi} S, 1 \right\}.$$

Verifying that the strategy described in the statement of Theorem 1 implements the upper bound is straightforward. This completes the proof.

Proposition 2

Sufficiency. Use the notation e_k for the binary K-vector for which $m_k = 1$ and $m_l = 0$ for all $l \neq k$. Consider the following strategy: When $\theta = Y$, play $m = (1, ..., 1)$ with probability one and randomize uniformly over all $D = \{k\}$, $k = 1, ..., K$. When $\theta = N$, randomize uniformly over all reports $(m, D) = (e_k, \{k\})$, $k = 1, ..., K$. It is easy to see that $\rho(m, D) = K$ for every (m, D) in the support of this strategy. Therefore, when $\pi \geq 1/(1 + K)$, the receiver always plays $a = y$. Let us now verify that the strategy is robust. First, by construction, the distribution over D is state-independent, thus satisfying part (i) in the definition of robustness. Second, given the message strategy, a type-H interpreter can attain his first-best with the following interpretation strategy: When $m = (1, ..., 1)$, he mimics the given interpretation strategy; and when $m = e_k$, he plays $D = \{k + 1 \bmod K\}$, thus inducing $a = n$ with the smallest possible dictionary.

Necessity. Suppose that σ is a robust strategy that attains full persuasion. Let $\mathcal{D}$ denote the set of all non-singleton dictionaries that are played with positive probability under σ. The proof will proceed stepwise, after making the following preliminary observation.

Observation 2 *Fix a message strategy $(\sigma(m \mid \theta))$ and consider two dictionaries D, D' such that $|D| \neq |D'|$. Then, for any realized m, neither sender type is indifferent between D and D'.*

This follows immediately from the lexicographic preferences.

Step 1: $\Pr(\mathcal{D}) < 1$.

Assume the contrary - i.e., no singleton dictionary is played in equilibrium. Consider a message realization m for which $\Pr(\theta = Y \mid m) < \frac{1}{2}$ under σ. Since $\pi < \frac{1}{2}$, there must exist such m. By the full-persuasion assumption, any D for which $\sigma(D \mid m) > 0$ satisfies $\rho(m, D) \geq (1 - \pi)/\pi$. By condition (ii) in the definition of robustness, it must be the case that

$$\rho(m, \{k\}) < \frac{1 - \pi}{\pi} \tag{9}$$

for every $k = 1, ..., K$ - otherwise, the type-A sender would use a singleton dictionary at m. It follows from (9) that a type-H interpreter would necessarily prefer to use a singleton dictionary at m. By condition (iii) in the definition of robustness, singleton dictionaries must be played with positive probability under σ, a contradiction. □

Step 2: *Suppose $|D| = 1$ for some report (m, D) that is played with positive probability under σ. Then, $|D'| = 1$ for every (m', D') that is played with positive probability under σ, such that $m'_D = m_D$.*
Assume the contrary - i.e. there exist reports (m, D) and (m', D') that are played with positive probability under σ, such that $|D| = 1$, $|D'| > 1$ and $m'_D = m_D$. By definition, $\rho(m', D) = \rho(m, D)$. Therefore, the realization (m', D') is inconsistent with condition (ii) in the definition of robustness. □

By Observation 2, we can partition the set of equilibrium messages into two classes: M_0 is the set of messages that are accompanied by singleton dictionaries, whereas M_1 is the set of messages that are accompanied by non-singleton dictionaries. Recall that

$$\Pr(m_D \mid \theta) = \sum_{(m',D') \mid m'_D = m_D} \sigma(m', D' \mid \theta)$$

By Step 2, if $m \in M_0$, the R.H.S summation only covers reports (m', D') such that $m' \in M_0$. Furthermore, by condition (i) in the definition of robustness, $\Pr(M_0 \mid \theta = Y) = \Pr(M_0 \mid \theta = N) = \alpha$ under σ. By Step 1, $\alpha > 0$.

It follows that we can rewrite the joint distribution over (θ, m, D) that is induced by σ as a three-stage lottery. In the first stage, *before* θ is realized, the classes M_0 and M_1 are drawn with probability α and $1 - \alpha$, respectively. In the second stage, θ is realized, where $\theta = Y$ with probability π, independently of the lottery's first stage. Finally, (m, D) is realized conditional on θ, with the restriction that m must belong to the class that was realized in the first stage.

Therefore, in order for the receiver to play $a = y$ with probability one, it must be the case in particular that he plays $a = y$ with probability one conditional on the realization M_0 in the first stage of the three-stage lottery. But this can only hold if the condition for full persuasion given in Remark 2 for the case of singleton dictionaries. Therefore, it must be the case that $\pi \geq 1/(1+K)$.

Proposition 3

Construct the following strategy for the sender.

Message strategy. In state Y, the sender randomizes uniformly between $m = (1, ...1)$ and $m = (0, ..., 0)$. In state N, he randomizes uniformly over the set of all messages m for which $m_k = 1$ for exactly $\lfloor K/2 \rfloor$ values of k.
Interpretation strategy. Every m that is played in state N is accompanied by $D = \{k \mid m_k = 1\}$. In state Y, the sender mixes uniformly over all sets D of size $\lfloor K/2 \rfloor$, independently of m.

By construction, $\Pr(m_D \mid \theta = Y) = \Pr(m_{D^c} \mid \theta = Y) = \frac{1}{2}$ and $\Pr(m_D \mid \theta = N) = \Pr(m_{D^c} \mid \theta = N) = 1/S$ for every (m, D) that is played. By (4), the receiver's likelihood ratio for any realized message m satisfies

$$\begin{aligned}\frac{\widetilde{\Pr}(m, D \mid \theta = Y)}{\widetilde{\Pr}(m, D \mid \theta = N)} &\geq \frac{\Pr(m_D \mid \theta = Y) \cdot \Pr(m_{D^c} \mid \theta = Y)}{\max\{\Pr(m_D \mid \theta = N), \Pr(m_{D^c} \mid \theta = N)\}} \\ &= \frac{\frac{1}{2} \cdot \frac{1}{2}}{\frac{1}{S}} = \frac{S}{4}\end{aligned}$$

The receiver will play $a = y$ whenever this expression is weakly above $(1-\pi)/\pi$.

Proposition 5

Denote $S(L) = \binom{L}{\lfloor L/2 \rfloor}$. Construct the message strategy first. In state Y, randomize uniformly over two messages: m^1 satisfies $m^1_k = 1$ for all $k \leq L$ and $m^1_k = 0$ for all $k > L$; m^2 satisfies $m^2_k = 0$ for all $k \leq L$ and $m^2_k = 1$ for all $k > L$. In state N, assign probability $\frac{1}{2}S(L)$ to every message m such that $m_k = 1$ for $\lfloor L/2 \rfloor$ values of $k \in \{1, ..., L\}$, and $m_k = 0$ for all other k. Likewise, assign probability $\frac{1}{2}S(L)$ to every message m such that $m_k = 1$ for $\lfloor L/2 \rfloor$ values of $k \in \{L+1, ..., 2L\}$, and $m_k = 0$ for all other k.

The conditional dictionary distribution is as follows. Conditional on any m that is played in state N, let $D = \{k \mid m_k = 1\}$ with certainty. Conditional on m^1, D

is distributed uniformly over all subsets of $\{L+1, ..., 2L\}$ of size $\lfloor L/2 \rfloor$. Finally, conditional on m^2, D is distributed uniformly over all subsets of $\{1, ..., L\}$ of size $\lfloor L/2 \rfloor$.

It is easy to verify that this strategy satisfies the two desiderata and induces the sender's first-best whenever $\pi \geq 1/(1+S(L))$.