Rules and Commitment in Communication: An Experimental Analysis

Guillaume R. Fréchette
NYU

Alessandro Lizzeri
NYU

Jacopo Perego
Columbia

September 27, 2019

Abstract

We investigate models of cheap talk, information disclosure, and Bayesian persuasion, in a unified experimental framework. Our umbrella design permits the analysis of models that share the same structure regarding preferences and information, but differ in two dimensions: the *rules* governing communication, which determine whether information is verifiable; and the sender's *commitment* power, which determines the extent to which she can commit to her communication strategy. Commitment is predicted to have opposite effects on information transmission, depending on whether information is verifiable. Our design exploits these variations to explicitly test for the role of rules and commitment in communication. Our experiments provide general support for the strategic rationale behind the role of commitment and, more specifically, for the Bayesian persuasion model of Kamenica and Gentzkow (2011). At the same time, we document significant quantitative deviations. Most notably, we find that rules matter in ways that are entirely unpredicted by the theory, suggesting a novel policy role for information verifiability.

1 Introduction

The goal of this paper is to experimentally study the effect of *commitment* on communication between a sender and a receiver who have conflicting interests. In any specific environment and application, knowing (and measuring) the exact degree of commitment available to an agent is difficult. However, the degree of commitment does vary depending on the context and commitment may depend on observable correlates such as the frequency of communication and the protocols governing such communication. Thus, understanding the effect of commitment on communication is a natural question. In this paper we present a model of partial commitment that generates comparative statics on the role of commitment. For a specific treatment with partial commitment, the model generates predictions about differences in behavior in two stages: a stage that takes place before the sender observes his private information and one that takes place after the sender observes his private information. These predictions directly speak to the ability of senders to take advantage of commitment as well as their willingness to undo such commitments when this is to their advantage. Variation across treatments in the degree of commitment generates additional predictions about the degree of communication as well as receivers ability to understand the effects of commitment. In order to further discipline our analysis, we also allow for a distinction between verifiable and unverifiable messages. Several predictions of the model are qualitatively different in the two contexts, and some comparative statics on the degree of commitment go in opposite directions.

For expositional purposes, we begin with an experimental analysis of a Bayesian persuasion game (Kamenica and Gentzkow (2011)). We assume that there are two (low and high) states, two (low and high) messages, and two (low and high) actions. The sender wants the receiver to choose a high action, whereas the receiver wishes to match the state. The prior is such that, without effective information transmission, the receiver would choose the low action. The sender has full commitment power in the selection of information structures. In equilibrium, the sender commits to sending the high message with probability one when the state is high and to randomize between the low and the high message when the state is low so as to make the receiver indifferent between choosing the low and the high action, thereby maximizing the ex-ante probability that the receiver chooses the high action. We find that, in the data, receivers behave in a way that is qualitatively consistent with the theory: they understand what messages should lead to higher posteriors (of the state being high) and are more likely to respond with the high action when they have a higher posterior. A crucial consequence of receiver behavior in our data is that it induces payoffs for the sender that reflect the key strategic tensions in the theory. That is, the best strategy for a sender faced with the receivers in the data is to partially hide information when the state is low (just as predicted in the model), but to be

slightly more truthful than predicted by the theory. Average behavior by senders in the data is not far from the predictions of the theory. In particular, the average strategy displays a degree of hiding of information that is close to the theoretical predictions. However, the average hides rich heterogeneity in behavior. We discuss this heterogeneity in detail later.

The analysis of this first treatment is useful as an initial exploratory analysis of commitment in communication. However, from this treatment alone, it is hard to evaluate the balance between the patterns of behavior that are in line with the predictions of the model and those that are not. More generally, it is difficult to draw general conclusions about the role of the key strategic forces within the model. First, to properly test whether or not senders make strategic use of commitment power, we should take into account their behavior in the counterfactual scenario where they cannot commit. Second, we would ideally like to test comparative statics that can only be explained by the strategic use of commitment power.[1] For these reasons, we propose a general structure that introduces partial commitment as well as allowing for verifiable information to generate a rich set of qualitative predictions that enables us to draw more meaningful conclusions about the effects of commitment.[2]

We model partial commitment as a probabilistic opportunity (that arises after the state is realized) to revise the choices that were made at the commitment stage. Specifically, before learning the true state, the sender publicly selects an information structure. Messages are sent to the receiver according to this information structure with probability ρ. This probability is common knowledge. We refer to this stage as the *commitment stage*. After observing the state, with probability $(1 - \rho)$, the sender can privately revise her choice of message. We refer to this stage as the *revision stage*. The higher ρ is, the higher the probability that the sender will not be able to revise her strategy, and hence the higher the extent to which she is committed to her initial information structure. In the limit case in which $\rho = 1$, the sender has full commitment and outcomes converge to the Bayesian persuasion model (Kamenica and Gentzkow (2011)) studied in our first treatment.

---

[1] For example, when information is unverifiable, if some of the senders are averse to lying, an increase in commitment power can increase the equilibrium informativeness. This change happens for reasons unrelated to the strategic use of commitment. Our framework allows us to conclude that this possibility cannot be the main factor in our data.

[2] A virtue of our experimental design is that it allows us to jointly analyze models that share an underlying structure regarding preferences and information, but that are distinguished either by the extent to which the sender can commit to her communication plan or by the rules governing communication, namely, whether information is verifiable or not. With minimal differences between treatments, our common structure encompasses models of cheap talk (Crawford and Sobel (1982)), models of disclosure (Grossman (1981), Milgrom (1981), Jovanovic (1982), Okuno-Fujiwara et al. (1990)), and models of Bayesian persuasion (Kamenica and Gentzkow (2011)); as well as intermediate cases between these extremes. Hence, we span a considerable portion of the models of strategic information revelation that have been discussed in the literature in the last decades, and we experimentally study novel dimensions of the sender-receiver interaction.

For any given level of partial commitment we further distinguish between verifiable and unverifiable information. This distinction is particularly useful because some key predictions of the model are qualitatively different in the two scenarios and some go in opposite directions depending on the verifiability of information. When senders' messages are *unverifiable*, senders can freely misreport their private information. When messages are *verifiable*, information cannot be misreported, but it can be hidden.[3]

Our first main finding in the data is that subjects understand the power of commitment: senders figure out how to exploit commitment and receivers figure out how to react to it. This can be seen by contrasting senders behavior in the commitment stage to their behavior in the revision stage. The theory predicts that in the case of unverifiable information, senders should reveal more information in the commitment stage than in the revision stage and that this ranking should be reversed when information is verifiable. The data strongly support this prediction of the theory. Our second main finding is that receivers understand that information conveyed in the commitment stage is more meaningful when the level of commitment is higher: in higher-commitment treatments, receivers are more responsive to information from the commitment stage. We then consider how commitment affects overall equilibrium informativeness, by comparing treatments with different levels of commitment. We find that subjects behave in ways that are consistent with the theory: informativeness *decreases* with commitment in the verifiable treatments and *increases* with commitment in the unverifiable treatments. Furthermore, we find that verifiability has the predicted effect of increasing the amount of information conveyed by senders. However, quantitatively, verifiability matters more than it should, so that informativeness does not rise enough with the degree of commitment in the unverifiable treatments and does not decrease enough in the verifiable treatments. This departure from the theory is therefore particularly visible in the limiting case of full commitment, $\rho = 1$: a lot more information is revealed when communication is verifiable than when it is not, despite the fact that, in theory, equilibrium informativeness is the same in the two treatments. This feature then drives us to investigate the full commitment cases in more detail. We discuss the extent to which models with boundedly rational or "behavioral" agents may help explain the patterns we find in the data. We estimate a Quantal Response Equilibrium model that allows us to compare full commitment treatments. We argue that the main reason why we observe these large differences in informativeness between

---

[3]The sender misreports her private information when she sends messages that are *false*. A message is false if none of the statements it contains are true. For example, provided that the ball is blue, message "the ball is red or black" is false. We will formalize this idea in Section 3.1. When there is no commitment, unverifiable information is associated with models of cheap talk (e.g., Crawford and Sobel (1982)), whereas verifiable information corresponds to models of disclosure (e.g., Grossman (1981), Milgrom (1981), Jovanovic (1982), Okuno-Fujiwara et al. (1990)).

verifiable and unverifiable treatments is that departures from equilibrium play are likely to *increase* the amount of information conveyed by senders under verifiable information and to *decrease* informativeness under unverifiable information. From a policy perspective, this excess informativeness under verifiable information presents a novel justification for increasing the difficulty for senders to misreport their information.[4] Finally, the theory predicts that senders convey more information when parameters are such that receivers need more evidence to go against their prior, and this result is consistent with what we find in the data, although, once again, this effect is quantitatively smaller than predicted by the theory.

We depart from the previous experimental literature on information transmission in several ways. First, we innovate by conducting an analysis *across* a variety of models. Of course, when performing such an exercise, ensuring that all sources of variations coming from seemingly unimportant details of the design are reduced to a minimum is crucial, so that differences in outcomes in the data can be imputed to differences in the treatments. In order to do so, we take advantage of our theoretical framework, thanks to which we are able to design an experiment that allows us to move from one model to another by simply changing one of the two parameters, namely, the degree of commitment on the sender's part and the verifiability of messages. An additional advantage of considering all these treatments under the same umbrella is that doing so provides discipline on the explanations that can be used to rationalize potential deviations from theoretical predictions.

A second way in which we depart from the previous experimental literature on cheap talk is that, rather than investigating the relationship between the informativeness of communication and the degree of preference alignment between the sender and the receiver, we focus on the effect of commitment.[5] Models with unverifiable communication have been used to study a variety of phenomena, including lobbying (Austen-Smith (1993), Battaglini (2002)), the relation between legislative committees and a legislature, as in, for example, Gilligan and Krehbiel (1989) or Gilligan and Krehbiel (1987), and the production of evidence to a jury (Kamenica and Gentzkow (2011), Alonso and Camara (2016)). Dranove and Jin (2010) survey the literature on product quality and the disclosure of information. A number of experimental papers study cheap talk. Blume et al. (2017) provides a survey of the experimental literature on communication. Dickhaut et al. (1995) is the first experimental paper to test the central prediction of Crawford and Sobel that more preference alignment between the sender and the receiver should result in more infor-

---

[4]We also find that, in the unverifiable treatments with a substantial amount of commitment, receivers are excessively skeptical. This finding is partly in contrast to prior literature on cheap talk (see the review by Blume et al. (2017) and the paper by Khalmetskia et al. (2017)).

[5]In doing so, our paper also addresses recent theoretical contributions on persuasion under partial commitment, e.g. Lipnowski et al. (2018) and Min (2017). In a similar spirit, Blume et al. (2019) also investigates experimentally changes to the communication environment, as opposed to preference misalignment.

mation transmission. Their main result is consistent with this prediction. Forsythe et al. (1999) add a cheap-talk communication stage to an adverse-selection environment with the feature that the theory predicts no trade and that communication does not help. By contrast, in the experiment, communication leads to additional trade, partly because receivers are too credulous. Blume et al. (1998) study a richer environment and compare behavior when messages have preassigned meanings with behavior when meaning needs to emerge. Among other findings, they confirm that, as in Forsythe et al. (1999), receivers are gullible. Cai and Wang (2006) find that senders are overly truthful and that receivers are overly trusting, relative to the predictions of the cheap-talk model. They also study information revelation as players' preferences become more aligned: consistent with the theory, they find the amount of information transmission increases with the degree of preference alignment. They then discuss how to reconcile the departures from the predictions of the cheap-talk model via a model of cognitive hierarchy and via Quantal Response Equilibrium.[6]

Models of disclosure of verifiable information have been used to study the disclosure of quality by a privately informed seller, for instance, via warranties,[7] of the contents of financial statements by a firm,[8] and in many other contexts. Dranove and Jin (2010) survey the literature on product quality and the disclosure of information. In contrast with experiments on cheap talk, experiments on the disclosure of verifiable information typically find under-revelation of information when compared with the theoretical predictions. For instance, Jin et al. (2016) find that receivers are insufficiently skeptical when senders do not provide any information, which in turn leads senders to underprovide information, thereby undermining the unraveling argument.[9] Some papers also study information unraveling with field data. In particular, Mathios (2000) studies the impact of a law requiring nutrition labels for salad dressings. He shows that, prior to mandatory disclosure, low-fat salad dressings posted labels, while a range of high-fat salad dressings chose not to disclose. Mandatory disclosure was followed by reductions in sales for the highest-fat dressings. These results are in conflict with the predictions of the unraveling result from the literature on verifiable communication. Jin and Leslie (2003) study the consequences of mandatory hygiene grade cards in restaurants. They show that hygiene cards lead to increases in hygiene scores, that demand becomes more responsive to hygiene, and that fewer food-borne-illness hospitalizations occur.

A third element of novelty in our design is the treatment under full commitment. As discussed above, this treatment coincides with a model of Bayesian persuasion as

---

[6]See also Sánchez-Pagés and Vorsatz (2007), Wang et al. (2010), and Wilson and Vespa (2017).

[7]For example, (Grossman (1981), Milgrom (1981), Jovanovic (1982), Okuno-Fujiwara et al. (1990)).

[8]See for instance, Verrecchia (1983), Dye (1985), and Galor (1985).

[9]See also Forsythe et al. (1989), King and Wallin (1991), Dickhaut et al. (2003), Forsythe et al. (1999), Benndorf et al. (2015), Hagenbach et al. (2014), and Hagenback and Perez-Richet (2018).

introduced in Kamenica and Gentzkow (2011). This model has become influential in the recent theoretical literature, which is comprehensively reviewed by Bergemann and Morris (2019).[10] Evaluating how the degree of commitment affects outcomes is one way to experimentally evaluate the model of Bayesian persuasion.

Our paper is one of three new experimental investigation of Kamenica and Gentzkow (2011). Nguyen (2017) and Au and Li (2018) both innovate with clever designs aimed at making the game easier for subjects to understand. Nguyen (2017) uses an intuitive interface for senders to enter their communication strategy. Furthermore, the communication strategy is discretized, and in the main experiment, the number of possible strategies the sender can use is small. Finally, given those simplifications, she can increase the number of repetitions to 80, allowing ample opportunities for learning. The experiment of Au and Li (2018) uses an implementation such that the sender can select posteriors directly, thus eliminating the need for receivers to do Bayesian updating. Other implementation differences are the use of a fixed partner design and a smaller number of repetitions with only 10 rounds. In addition, they consider the predictions of a modified model in which preferences are such that agents have other-regarding concerns. Finally, they test a specific prediction of that model by considering two treatments that vary the prior. Both experiments find that senders, on average, and as predicted, convey less than full information. In particular, Nguyen (2017), who has the simplest setting and more repetitions, finds that a high fraction of senders behave optimally, given receivers' behavior, and that their behavior involves hiding some information. They both report that receivers are more likely to go against their prior as their posterior increases. In addition, they also both find that when the posterior on the state the sender prefers is at 0.5, the likelihood of a receiver guessing in a way that benefits the sender is far from certain (in both studies around 50%). These results are also consistent with our findings, which suggests these results are robust given that all three implementations are fairly different.

2 Benchmark Treatment of Bayesian Persuasion

2.1 The Game and its Implementation

The Game. In our baseline treatment, we implement the following sender-receiver game. A ball is drawn from an urn containing three balls: Two are blue (B) and one is red (R). The color of a ball represents the realization of a payoff state, which we denote $\theta \in \{B, R\}$. The prior probability that the state is R is $\mu_0(\theta = R) = \frac{1}{3}$. The first stage of the game is a commitment stage: The sender commits to an *in-*

---

[10]Some recent papers includeGentzkow and Kamenica (2014), Bergemann et al. (2015), Alonso and Camara (2016), Duffie et al. (2017), Bardhi and Guo (2018), Galperti (2019)

formation structure, namely, a map from states to (possibly random) messages. In this treatment, we allow the sender to choose among two messages, denoted r and b. The second stage of this game is a guessing stage: The receiver observes the information structure as well as a message generated by the information structure. Her task is to make a guess $a \in \{red, blue\}$. Players' preferences are described in Table 1: The receiver wants to correctly guess the state, while the sender would like the receiver to always guess $a = red$, irrespective of the state.[11]

Table 1: Payoffs

Guess (a)	State (θ)			
	R		B	
red	Receiver $2	Sender $2	Receiver $0	Sender $2
blue	Receiver $0	Sender $0	Receiver $2	Sender $0

Equilibria. This game has several payoff-equivalent Perfect Bayesian equilibria with a common structure. In this section, we focus on the following equilibrium featuring "natural language:" conditional on state $\theta = R$, the sender commits to sending message r with probability 1; conditional on state $\theta = B$, she commits to sending messages r and b with equal probability.[12] This information structure maximizes the ex-ante probability that the receiver guesses *red*: it induces a posterior of zero following message b and a posterior of $1/2$ following a message r. Thus, the receiver guesses *blue* following a message b and is willing to guess *red* following message r because she is indifferent between *red* and *blue*.[13] Two simple features of equilibrium stand out. First, the sender benefits from commitment. In this setting commitment allows credible communication. When based exclusively on her prior information (also the equilibrium outcome of the game without commitment) the receiver's guess would always be $a = blue$. By committing to an appropriate information structure, instead, the sender can persuade the receiver to guess $a = red$, at least some of the time. Second, the sender's optimal communication strategy involves *partial* information revelation requiring randomization among messages conditional on state B.

---

[11]Note that one of the advantages of our design is that predictions are independent of risk preferences because outcomes are binary.

[12]As we illustrate shortly, the use of natural language is indeed predominant in our data.

[13]In equilibrium, the receiver must choose *red* with probability 1 following message r because otherwise the sender would choose an information structure that induces a slightly higher posterior conditional on message r, but then the sender would have no best response.

7

Implementation in the laboratory.[14] At the beginning of each session, instructions were read aloud, and subjects were assigned a fixed role (sender or receiver). In each session, subjects played 25 paid rounds of the game described above with random rematching between rounds. We conducted four sessions lasting approximately 100 minutes each. Sessions included 14-20 subjects (17.5 on average per session) for a total of 70 subjects. In addition to their earnings from the experiment, subjects received a $10 show-up fee. Average earnings, including the show-up fee, were $34 (ranging from $14 to $52) per session.

In our experiment, a key feature is the choice of information structure by the sender. Our design makes this choice particularly straightforward and easy to visualize. Senders simply move sliders on the screen, and the color of each bar reflects the chosen probabilities for each message as displayed in Figure B16 of Appendix B. These probabilities are updated in real-time in the cells above the sliders. The receiver observes the information structure chosen by the sender and makes a guess for each possible message (strategy method). The specific choice probabilities for each message can be seen by dragging the mouse cursor over the communication strategy. Appendices B and C contain a sample of the instructions and more detailed information on the implementation in the laboratory.

The results reported in this and the next sections are computed using the data from the last 10 rounds of play in each session. We discard earlier rounds to allow enough time for subjects to familiarize themselves with the experiment and to learn the relevant strategic forces in the task they are facing.[15]

Measuring Informativeness. The "amount" of information the sender transfers to the receiver, namely, the *informativeness* of her communication strategy, represents a variable of central interest in our analysis. We will report different measures of informativeness, because they present different advantages, some being easier to interpret, others allowing for more disaggregated analysis, etc.

Our main measure of informativeness is the correlation coefficient between the color of the ball and the receiver's guess.[16] We denote this variable by ϕ. To fix ideas, suppose the sender truthfully discloses the color of the ball. Then, the receiver's final guess should be perfectly correlated with the state. Conversely, if the sender babbles, the receiver's final guess will be uncorrelated with the the state.[17]

---

[14]Subjects were recruited from the NYU undergraduate population using hroot (Bock et al., 2014).

[15]Appendix D reports some results on how subjects' behavior evolves over rounds for the entire experiment.

[16]This measure has been extensively used in the experimental literature on communication. See, for instance, Forsythe et al. (1999), Cai and Wang (2006), and Wang et al. (2010).

[17]Our design allows us to leverage the power of the strategy method to obtain significantly more precise measures of ϕ and ϕ^B. In fact, we observe the complete strategies of both senders and receivers and, therefore, we can analytically compute the Pearson correlation coefficients (specifically, the phi coefficients, since our variables are binary). This is as if we could observe an infinite sample

This way of measuring informativeness has the potential drawback of compounding the mistakes of both senders and receivers. Suppose for instance that the sender truthfully discloses the state, but the receiver does not listen. In this case $\phi = 0$, although a great deal of information was offered to the receiver. To isolate the sender's behavior from the mistakes of the receivers, we will use an alternative measure of informativeness, the *Bayesian* correlation, denoted ϕ^B. This is the correlation coefficient implied by the sender's strategy combined with the guesses of a hypothetical *Bayesian* receiver.

The Bayesian correlation coefficient, however, hides potentially useful information. For example, a sender who generates a posterior conditional on message r that is just below 0.5 does convey some information to the receiver. Nonetheless, this posterior leads to a correlation of zero because the Bayesian receiver would choose *blue* in both states following such a posterior; that is, the same correlation as if the sender had conveyed no information (a posterior of 1/3). Hence, it will sometimes be useful to directly consider the distrubution of induced Bayesian posteriors.

2.2 Results for the Benchmark Treatment

We now present a number of facts that help us characterize the behavior of senders and receivers in this treatment. We begin with a description of receiver behavior. We then discuss the consequences of this behavior for senders' payoffs, and then proceed with an analysis of sender behavior.

2.2.1 Receivers' Behavior

In our discussion of receivers' behavior, we take as given the information conveyed by senders' strategies that we discuss in more detail below. Our first objective is to understand how receivers respond to such strategies.[18] Receiver's responses in turn generate the payoffs that senders face when selecting their strategies.

We begin our analysis of receivers by describing some key aggregate features of the data. On average, receivers react to a higher posterior $\mu(m)$ by guessing *red* with higher frequency, as illustrated in Figure 1. Thus, receivers present monotonic behavior: they are more persuaded to guess *red* by messages that carry more evidence in favor of the state being R. For instance, for posteriors above $\frac{1}{2}$, receivers guess *red* 57% of the time, whereas they guess *red* only 11% of the time for posteriors below $\frac{1}{2}$ ($p \leq 0.01$).[19]

---

of *realized* states and guesses per each round. Simulations we have done suggest the improvement in precision from using our method is non-trivial and that samples would otherwise need to be large for the estimates of the Pearson correlation to stabilize.

[18] In our environment, receivers make guesses in a relatively straightforward setting. In contrast, Epstein and Halevy (2019) analyze receiver's behavior when the setting in more complex.

[19] Unless noted otherwise, all statistical results allow for random-effects at the subject level and
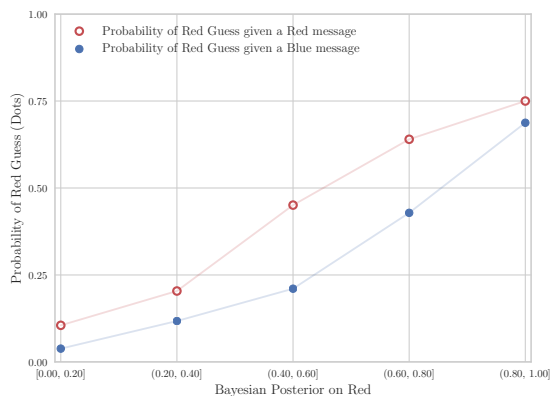
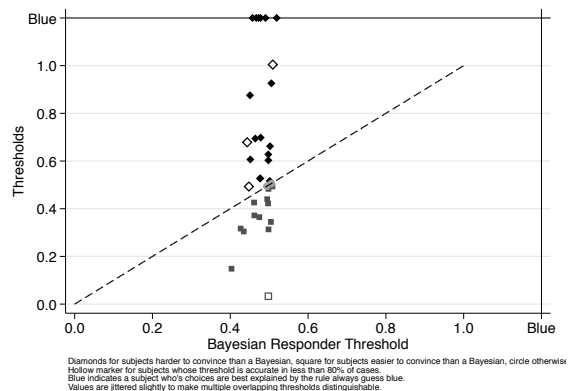Figure 1: Probability of Guessing Red by Posterior and Message

Figure 2: Estimated Thresholds: Actual Receivers vs Bayesians

The monotonicity displayed in Figure 1 is, of course, a mild requirement for receivers' rationality: given the payoffs in our experiment, a Bayesian receiver should respond by guessing *red* with probability one for any posterior $\mu(m) \geq \frac{1}{2}$, and by guessing *blue* with probability one otherwise. Clearly, the aggregate evidence from Figure 1 fails to fulfill this stronger requirement of rationality. Furthermore, receivers respond to the color of the message independently of the posterior this color conveys. When $\mu(m = r) \geq \frac{1}{2}$, receivers guess *red* 62% of the time following an r message and 38% of the time following a b message. In contrast, when $\mu(m = b) < \frac{1}{2}$, receivers guess *red* 21% of the time following a r message and 5% of the time given a b message. These differences, which are significant at the 1% level, are inconsistent with the behavior of a Bayesian receiver. Even when provided with conclusive evidence that the state is R, that is, even when $\mu(m)$ is very close to 1, some receivers nonetheless guess *blue*, at least some of the time. To summarize, aggregate receivers' behavior does not correspond exactly to the Bayesian paradigm, an observation in line with other experiments that documents non-Bayesian behavior of subjects in laboratory experiments (see, e.g., Charness and Levin (2005) and Chapter 30 of Holt (2007) for an overview). Nonetheless, behavior in aggregate does react in the direction of Bayesian behavior (monotonicity has been documented in other experiments; see Camerer (1998) for a discussion). To understand better whether the deviations are driven by a few subjects or shared by most, we turn to individual behavior.

We now demonstrate that there are systematic patterns in how receivers react to the information they receive, as summarized by the posterior belief. In particular, we consider the possibility that subjects follow (potentially different) *threshold strategies*, that specify guessing *red* if and only if their posterior is weakly above

---

are clustered at the session level. We include random-effects to account for persistent heterogeneity across subjects; clustering is motivated by potential session-effects (see Fréchette, 2012). Results for alternative specifications are reported in the appendix. We note that the findings in the alternative specifications suggest that session-effects are not important in this setting.

a certain threshold $\bar{\mu}$. For example, if $\bar{\mu} = \frac{1}{2}$, the receiver is, indeed, Bayesian. If $\bar{\mu} > \frac{2}{3}$, instead, the receiver is not Bayesian and yet her behavior is systematic and can be said to require stronger evidence to choose *red* than a Bayesian would. We now estimate the receiver specific threshold that rationalizes the greatest fraction of her guesses.[20] We find that the behavior of many subjects is consistent with a threshold rule. Almost half the receivers (46%) display behavior that is always consistent with a threshold strategy, and almost nine out of ten receivers (89%) are consistent with a threshold strategy for more than 80% of their guesses. Figure 2 plots the estimated threshold for each receiver as a function of the threshold that we would have estimated from the same data if that particular receiver were Bayesian.[21] As the figure shows, substantial heterogeneity in receivers' behavior exists. Dots lying above the 45-degree line indicate receivers who are reluctant to guess *red*, even when a Bayesian would conclude that there is enough evidence. By contrast, the points below the 45-degree line indicate subjects who are too eager to guess *red*, despite insufficient evidence from the perspective of a Bayesian. The aggregation of this heterogeneous behavior is partly responsible for the smoothness of aggregate responses to the posterior that is displayed in Figure 1. Also note that Figure 2 shows a sizable fraction of receivers who exhibit behavior consistent with the Bayesian benchmark: One quarter of the receivers have thresholds within five percentage points of being consistent with a Bayesian receiver; the number increases to one third if we are more permissive and allow for a band of ten percentage points around the Bayesian receiver.

2.2.2   A Best-Response Analysis

We now wish to understand the consequences of receiver behavior for the sender incentives. In particular, it is important to know whether the key strategic tension that is present in the Bayesian persuasion model survives in the world of imperfect rationality represented by the receivers in our data. To answer this questions, we first fit a probit model to estimate how each receiver would map any given posterior into a probability of guessing *red*. Given this estimated model of receivers' behavior, we consider an important class of senders' strategies and we compute their hypothetical expected payoff. More specifically, we define a class of strategies that can be parametrized by a single parameter, but that is rich enough to accommo-

---

[20]Because we focus on the last 10 rounds of the game, and because we use the strategy method for the receivers, we observe a receiver's guess on 20 occasions following r and b messages. We look for the threshold that best describes these 20 observations. This procedure typically results in a *range* of best-fitting thresholds, of which we report the average one. See Appendix D.1 for a more detailed explanation.

[21]Given finite data, even a Bayesian receiver can have an estimated average threshold that is different from $\frac{1}{2}$. As an example, imagine a receiver who is perfectly Bayesian, but for whom the closest posteriors to 0.5 that we observe were 0.45 and 0.65. Her estimated threshold would then be 0.55. Figure D21 in the appendix presents the estimated threshold and their respective precision.
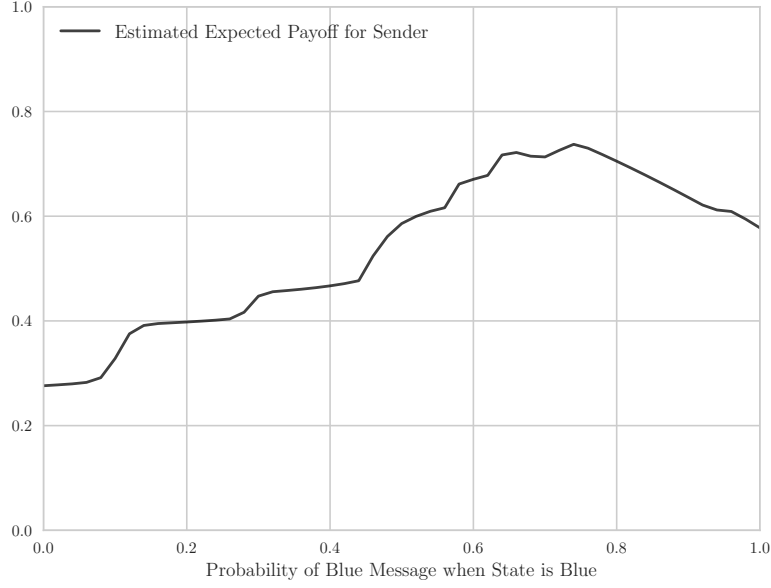
Figure 3: Empirical Expected Payoffs

date almost all the strategies that are actually chosen by senders in the data (as we will see later in Figure 6): if $\theta = R$, the strategy sends message r with probability one, if $\theta = B$, the strategy sends message r with probability γ and message b with probability $1 - \gamma$. Therefore, γ parametrizes the extent of informativeness of these strategies.[22] Figure 3 displays the expected payoff for these strategies, as a function the parameter γ. Figure 3 confirms an important qualitative insight from the theory of Bayesian persuasion. The senders' expected payoff is non-monotonic in the amount of information conveyed to the receiver. In our data, as in the theory, being completely uninformative is worse than being entirely truthful, which is, in turn, worse than engaging in some degree of strategic mixing. However, the best-response consists of overshooting a bit, that is, providing more information than required by the equilibrium benchmark. Receivers' departures from Bayesian behavior also lead to a payoff function for senders that is flatter and smoother than if senders faced a population of Bayesian receivers. This feature can potentially make it more difficult for senders to learn to play along the lines predicted by the theory.

2.2.3 Sender Behavior: Types and Informativeness

We now turn to our analysis of sender behavior. This description is challenging as the game is complex and strategies are high-dimensional objects. We proceed in several steps, starting at an aggregate level, and then moving to a less aggregated one.

---

[22]For example, when $\gamma = 1$, the strategy is entirely uninformative, and all its induced posteriors are equal to the prior; when $\gamma = 0$, the strategy is perfectly informative and the induced posteriors are either 1, for the red message, or 0, for the blue message. Finally, the equilibrium strategy specifies $\gamma = 0.5$.
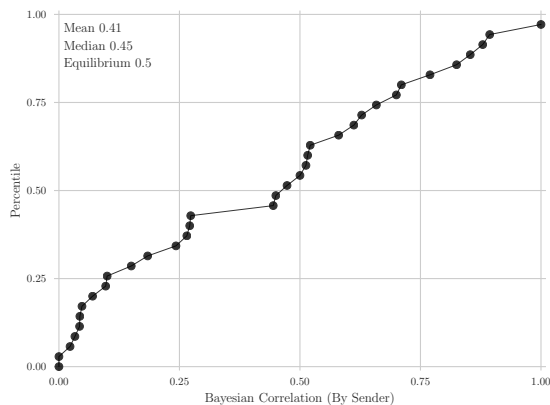
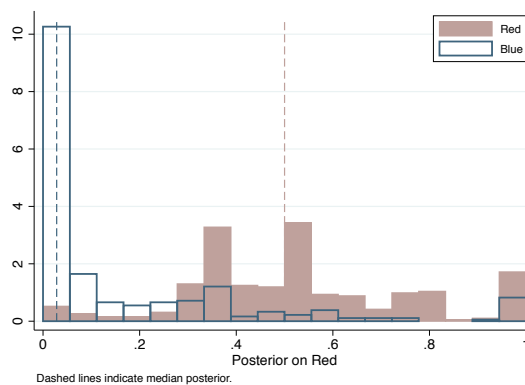Figure 4: CDF of Subject Average Bayes Correlation

Figure 5: Histogram of Posterior on Red by Message

We begin by studying the informativeness of sender behavior. In this benchmark treatment, the correlation coefficient implied by the equilibrium strategies is 0.50.

Figure 4 plots the distribution of the sender-specific average Bayesian correlations. The overall (i.e. across-sender) average correlation is 0.41 and the median is 0.45. Although this average is fairly close to the equilibrium value of 0.50, the distribution in Figure 4 clearly shows a high level of heterogeneity: some senders rarely reveal any information, and others consistently reveal almost all the information. However, a non-negligible group of subjects conveys some, but not all, of the information, as predicted in equilibrium.

We look at sender behavior in more detail by considering the empirical distribution of Bayesian posteriors that are induced by the observed senders' strategies in Figure 5.[23] The figure reveals a few important facts. First, consistent with equilibrium predictions, a blue message predominantly carries conclusive evidence that the state is $\theta = B$. Indeed, the most common posterior conditional on a blue message is close to zero. By contrast, the posteriors conditional on a red message are for the most part far from zero, but also highly dispersed. The most common posterior is close to 0.5, in line with the equilibrium prediction. The other spikes in this distribution of posteriors, at $1/3$ and at 1, represent clusters of strategies that we discuss next.

Many possible strategies can generate a particular amount of informativeness. We now describe the strategies actually chosen by the senders. We aggregate the data into groups of similar strategies using a k-means clustering analysis of senders' probabilities of sending each message as a function of the state.[24] The results indi-

---

[23]Because the state θ is binary, posteriors can be cast into the unit interval. As a convention, the *posterior* is the conditional probability of the state being R, that is, $\mu(m) := \mu_0(\theta = R|m)$.

[24]K-means clustering (MacQueen, 1967) is a commonly used method to group data— a form of unsupervised learning—see Hastie et al. (2009) and Murphy (2012) for a recent treatment. The procedure selects points to be the centers of clusters: a point is associated with the closest center, and the centers are iterated on to minimize the total within cluster variance.

cate that almost 90% of the observed choices can be organized into three *clusters*, whose representative strategies, or *types*, are displayed in Figure 6.[25] These strategies share one common feature: On average, the probability of sending message r conditional on state $\theta = R$ is close to 95% (median 99%), consistent with the equilibrium prediction. However, these strategies differ substantially in the probability with which the sender reports message b conditional on state $\theta = B$. For the three types, this number is 89%, 52%, and 10%, respectively. We compute the average correlation coefficient with Bayesian receivers, that is, ϕ^B (defined above), for each cluster and we find values of 0.82, 0.35, and 0.03, respectively (median values are 0.81, 0.50, and 0.00, respectively). The implication is that the three clusters identify three substantially different *styles* of communication. The first one (representing approximately 23% of the data) is particularly truthful and reveals a lot of information. The last one is uninformative (24% of the data). Finally, the intermediate and most prevalent cluster (35% of the data), is qualitatively in line with the equilibrium prediction, both in terms of the induced correlation and in terms of the type of strategy chosen by the senders.
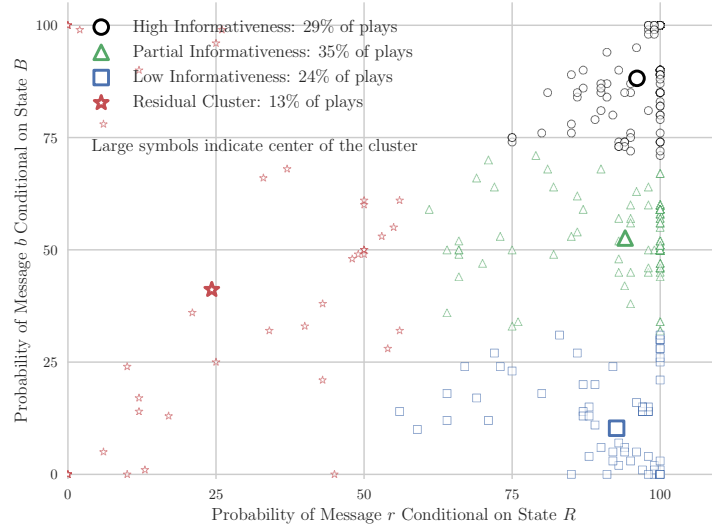

Figure 6: Sender's Strategies Grouped in Clusters

Importantly, the clustering analysis identifies types that are persistent over time. That is, our analysis illustrates that over rounds, senders in a cluster tend to play strategies from within the same cluster. For example, the median sender plays a

---

[25]The remaining 13% of observations fit into a "residual cluster" that is more difficult to recognize. In this exercise, we use clusters for descriptive purposes. Therefore, determining the best number of clusters is not of great importance. However, one method for selecting the number of groups, the elbow method, would select four groups. In addition, we note that our results are robust to using two different methods to determine starting values for the algorithm: initialized by using random groups or using the output of a clustering exercise on the Bayesian correlation (ϕ^B) as starting groups.

strategy that belongs to the same cluster nine times out of ten (more details in Figure D30).

2.3   Summary

In conclusion, this section analyzed a simple implementation of a Bayesian persuasion game and uncovered a set of basic properties that characterize the behavior of senders and receivers in the laboratory. The main conclusions that emerge from the analysis are the following. Although senders and receivers behavior is heterogeneous, the vast majority of subjects behave in systematic ways that are easily interpretable. Most senders engage in communication strategies that are sophisticated and do so by employing a natural language, but they differ in the amount of information they are willing (or able) to transfer to the receivers. Receivers react to information in a predictable manner; they understand the basic content of different information structures, but they differ in the amount of information they require in order to be persuaded to guess red. Finally, our analysis finds that a sizable group of subjects conforms with behavior that is consistent with the central qualitative insight that emerges from the theory of Bayesian persuasion. In particular, the sender is predicted to engage in some extent of strategic lying. According to the theory, neither full disclosure nor babbling is optimal for the sender: the sender should lie just enough to persuade. Figure 3 shows that aggregate receiver behavior generates payoffs for the sender that are consistent with this central qualitative insight, in spite of the fact that receivers' behavior is heterogeneous and sometimes far from Bayesian. Our discussion of senders' behavior shows that a non-negligible fraction of senders does respond to these incentives in a manner that is consistent with the theory: misrepresenting the state some, but not all, of the time when it when to their advantage.

3   The General Framework

In this section, we introduce a model of communication that is richer than the one of Section 2, and we discuss its main predictions. Later, we describe the experimental design and the equilibrium outcomes that obtain under the specific parametrization of our model that we bring to the laboratory. The model we present in this section differs from the benchmark treatment discussed before in two ways. First, we weaken the commitment assumption and allow for *partial* commitment. Second, we introduce two distinct communication rules: we contrast scenarios with *verifiable* and *unverifiable* information.

3.1 Theory

Let $\Theta = \{\theta_L, \theta_H\}$ be the state space and $\mu_0 \in [0,1]$ denote the common prior probability that the state is θ_H. There are two players: a sender and a receiver. The sender has private information about the state, the receiver has the ability to act. Communication consists of the sender transmitting information in an attempt to influence the action chosen by the receiver. The receiver chooses actions in a binary set $A = \{a_L, a_H\}$ and her preferences are given by the following utility function:

$$u(a_L, \theta_L) = u(a_H, \theta_H) = 0, \quad u(a_L, \theta_H) = -(1-q), \quad u(a_H, \theta_L) = -q.$$

Thus, the receiver wishes to match the actions to the state, and the relative cost of the mistakes in the two states is parametrized by q. A Bayesian receiver would choose action a_H whenever her posterior belief that the state is θ_H is larger than q. Thus, we call q the *persuasion threshold*.

The sender's preferences are state-independent and given by $v(a) := \mathbb{I}(a = a_H)$. That is, the sender receives a positive payoff only if she successfully persuades the receiver to take action a_H. We assume that $\mu_0 < q$. That is, absent further information, the receiver would choose a_L. Without this assumption, the sender would have no reason to communicate with the receiver and their interaction would be uninteresting. The sender communicates with the receiver by sending information about the state. An information structure is a map $\pi : \Theta \to \Delta(M)$, with $M = \{\theta_L, \theta_H, n\}$ representing the set of possible messages. We denote by Π^U the set of *all* information structures.

We consider different *communication rules*. A rule is an exogenous restriction imposed on the set of information structures $\Pi \subseteq \Pi^U$ that the sender can use when communicating with the receiver. We say that information is *unverifiable* if no restrictions are imposed on the sender, namely if $\Pi = \Pi^U$. We say that information is *simple* if the sender is restricted to binary messages, formally $\Pi = \Pi^S := \{\pi \in \Pi^U : \forall \theta, \ \pi(n|\theta) = 0\}$. We say that information is *verifiable* if, instead, $\Pi = \Pi^V := \{\pi \in \Pi^U : \pi(\theta_H|\theta_L) = \pi(\theta_L|\theta_H) = 0\}$. Under verifiable information, conditional on state θ, only messages in $\{\theta, n\}$ can receive positive probability. Therefore, we can interpret message $\theta \in M$ as a certifiable statement asserting that that the state is θ. Conversely, we can interpret message $n \in M$ as a statement that is neither true nor false and, hence, cannot be verified. In other words, verifiability demands that message θ can only be sent by type θ. Therefore, it is natural to assume that the receiver's belief upon observing message θ_H (resp. θ_L) is 1 (resp. 0), even if such a message occurs with zero probability.[26]

The game unfolds in three consecutive stages. First, in the *commitment stage* the sender chooses a strategy $\pi_C \in \Pi$ before learning the state θ. That is, the sender

---

[26] See Battigalli and Siniscalchi (2002).

commits to releasing information contingent on a state that she doesn't yet know. Second, in the *revision stage*, at every history (π'_C, θ), the sender chooses a strategy $\pi'_R \in \Pi$. Note that, π'_R can explicitly depend on π'_C, but we omit this dependence for notational simplicity. Moreover, it is important to remark that, because the sender learns θ before choosing π_R, she no longer has commitment power. Finally, in the *guessing stage*, for every history (π'_C, π'_R, m), the receiver observes (π'_C, m) and takes an action $a(\pi'_C, m) \in A$. It is commonly known that message m realizes with probability $\rho \in [0, 1]$ from π'_C and $(1 - \rho)$ from π'_R. The receiver updates beliefs according to some belief assessment $\mu(m, \pi'_C, \pi_R)$, an equilibrium object, that assigns a posterior belief to each message m, possibly as a function of π'_C and π'_R.

We refer to ρ as the sender's *degree of commitment*. It measures the extent to which sender is able to commit to her initial strategy π_C. For high values of ρ, the commitment strategy π_C is likely to be the one that determines the final message m. Conversely, for low values of ρ the final message m is likely to be determined by the choice at the revision stage, after the sender has learned the state.[27] Summing up, our framework is characterized by three main features that we shall exploit in the experiment: the communication rules Π, the degree of commitment ρ and the persuasion threshold q.

Conveniently, our framework nests several classic communication models as special cases. When $\rho = 0$ and information is unverifiable, our model captures cheaptalk communication (Crawford and Sobel (1982)). When $\rho = 0$ and information is verifiable, our model captures a disclosure game with verifiable communication (Grossman (1981), Milgrom (1981), Jovanovic (1982), Okuno-Fujiwara et al. (1990)). Finally, when $\rho = 1$ and information is unverifiable, our model becomes a Bayesian persuasion game (Kamenica and Gentzkow (2011)).

We use Perfect Bayesian Equilibrium (PBE) as a solution concept. In some cases, it is useful to focus attention on equilibrium outcomes, rather than equilibrium strategies. An outcome of particular interest in a communication game is how informative an equilibrium is: the extent to which the sender successfully communicates with the receiver. We measure informativeness of an information structure as the correlation between the state and the guess of a Bayesian receiver. Formally, fix $\pi \in \Pi$ and define $a^B(m, \pi) \in A$ to be equal to a_H if and only if $\mu(m, \pi) \geq q$. We define the *informativeness* of π to be $\phi^B(\pi) := \mathrm{Corr}(\theta, a^B(m, \pi))$. Note that $\phi^B(\pi)$ is non-negative and it is unaffected by how a^B is defined at messages that have zero probability under π.[28] We say that an information structure π is *more informative*

---

[27]Equivalently, one can think of the sender as having an *opportunity* to revise her commitment strategy after learning the state, which occurs only with probability $1 - \rho$. An alternative interpretation of the game is that the revision game is always available but the sender has a type that determines whether she will take advantage of the opportunity to revise the strategy. The parameter ρ is then the probability that the sender is not this opportunistic type.

[28]Informativeness can be measured in different ways, in particular in ways that do not depend on

than π' if $\phi^B(\pi) \geq \phi^B(\pi')$; an information structure π is *uninformative* if $\phi^B(\pi) = 0$, whereas it is *fully informative* if $\phi^B(\pi) = 1$. These definitions naturally extend to equilibria. More specifically, we say that equilibrium (π_C, π_R, a, μ) under (Π, ρ, q) is more informative than equilibrium $(\pi'_C, \pi'_R, a', \mu')$ under (Π', ρ', q') if $\rho\pi_C + (1-\rho)\pi_R$ is more informative than $\rho'\pi'_C + (1 - \rho')\pi'_R$.

As in many other communication games, our framework allows for multiple PBEs. For the interested reader, we provide a full equilibrium characterization of the equilibrium in Appendix A. In the rest of the paper, instead, we impose a simple tie-breaking rule on equilibrium behavior that is inspired by Hart et al. (2017). We say that a PBE is *truth-leaning* if, whenever it is optimal for type θ_H in the revision stage to tell the truth, she prefers to do so. A few comments are in order. First, this tie-breaking rule is simple but powerful. As we show in Proposition 1, it is sufficient to guarantee uniqueness of the equilibrium outcomes. Second, it is weaker than the refinement introduced in Hart et al. (2017). In fact, it is not imposed on all types of sender, but only on type θ_H, for which lying when indifferent is, perhaps, even less compelling. A fortiori, our tie-breaking rule is consistent with most of the equilibrium refinement that have been proposed in the literature. Third, this tie-breaking rule captures behavior that we do observe in our data.[29] In the rest of our analysis, we maintain the specialization to truth-leaning PBEs and we refer to these, more simply, as "equilibria," without further qualification.

3.1.1 Comparative Statics

Our goal is to develop a set of comparative statics for our framework, which we later use as experimental tests for the role of commitment in communication. We being with a characterization of equilibrium informativeness for a given level of commitment power ρ. This result also provides a contrast between the equilibrium informativeness at the commitment and at the revision stage.

Proposition 1. *Fix* $\rho \in [0, 1]$ *and* $q > \mu_0$. *Let* $\underline{\rho} := \frac{q - \mu_0}{q(1 - \mu_0)}$ *and* $\bar{\rho} := \frac{q(1-\mu_0)}{q(1-\mu_0)+(1-q)\mu_0}$ *and note that* $\underline{\rho} \leq \bar{\rho}$:

> [Unverifiable Information] *All equilibria are equally informative. Moreover, equilibria are* uninformative *if and only if* $\rho < \underline{\rho}$. *When* $\rho \geq \underline{\rho}$, *less information is transmitted at the revision stage than at the commitment stage.*

---

u. Our principal focus is on ϕ^B, in line with existing experimental literature on communication (e.g. Cai and Wang (2006)). Nonetheless, in our data analysis, we do consider alternative measures of informativeness that are based on the variance of the induced posterior beliefs, either unconditional or conditional on the state θ. These alternative measures of informativeness do not change the qualitative conclusions of our analysis, but are useful to highlight different aspects of the phenomena of interest. For more discussion, see Section 4.2 and Appendix D.

[29]For example, when information is verifiable, the average $\pi_R(\theta_H|\theta_H)$ in our data is about 0.95.

[Verifiable Information] *All equilibria are equally informative. Moreover, equilibria are* fully informative *if and only if $\rho < \bar{\rho}$. When $\rho \geq \bar{\rho}$, more information is transmitted at the revision stage than in the commitment stage.*

This result establishes uniqueness of the equilibrium outcomes and highlights the main tension between commitment and revision stages. It also emphasizes that this tension manifests itself in opposite ways under the different verifiability scenarios, thus providing a useful and easily testable prediction that we will exploit in our experimental analysis. To understand this result, we first consider the extreme cases: When $\rho = 0$, the sender has no commitment power. Therefore, equilibria are fully informative when information is verifiable and uninformative otherwise. When $\rho = 1$, instead, the sender has full commitment power. The equilibria feature partial information revelation in both the verifiability scenarios that we consider. The intuition for Proposition 1 is then the following. Under both verifiable and unverifiable information, the sender would like to commit to persuading the receiver to choose the high action as often as possible, and this requires partial information revelation. However, at the revision stage, a rational sender is unable to resist the temptation to undo her commitments and manipulate information in her favor. Under verifiable information, this opportunity implies full information disclosure in the revision stage; under unverifiable information, it implies sending the message that induces the high action, regardless of the state. The presence of this revision stage changes the sender's problem in the commitment stage relative to the full commitment scenario as follows: the sender over-communicates when information is unverifiable and under-communicates when information is verifiable. This modification is an attempt to obtain a final posterior for the receiver which is as close as possible to the full commitment scenario. When ρ is sufficiently high, partial information revelation occurs in both verifiability scenarios. This is because the revision stage cannot completely undo the positive effect of the commitment stage. Overall, this result illustrates how changes in the rules of communication can generate stark contrasts in they way senders react to commitment power.

Our next result describes how equilibrium informativeness changes with commitment power, and how this change depends on the rules of communication.

Proposition 2. *Fix $q > \mu_0$. When information is unverifiable, equilibrium informativeness weakly increases in ρ. When information is verifiable, equilibrium informativeness weakly decreases in ρ. Moreover, when $\rho = 1$, equilibrium informativeness is independent of the rules of communication.*

This result provides a clear set of empirical predictions suggesting experimental treatments to evaluate commitment. It illustrates how changes in the rules of communication can generate stark contrasts in the prediction of our model, allowing for a strong test of the role of commitment. The intuition for this result follows from

the discussion above. As ρ increases, the revision stage becomes increasingly less likely, and the relevance of the commitment stage increases. This increase allows the sender to approach the optimal solution under full commitment, $\rho = 1$. In our game, the solution for $\rho = 1$ is independent of the rules of communication, therefore leading to the same equilibrium informativeness. Under full commitment and verifiability, the sender can replicate the strategy that she uses when information is unverifiable, by replacing the use of message θ_H with message n.

Propositions 1 and 2 constitute the bulk of our experimental strategy to test the role of commitment, which revolves around the idea of partial commitment. Later in Section 4.4, we consider a different kind of comparative statics result that keeps fixed the degree of commitment ρ and shows how equilibrium informativeness changes with the persuasion threshold q.

Proposition 3. *Fix $q' > q > \mu_0$ and consider any $\rho \geq \frac{q' - \mu_0}{q'(1 - \mu_0)}$. Equilibrium informativeness under q' is strictly higher than under q, irrespective of the rules of communication.*

This result shows that, when the degree of commitment is sufficiently high, an increase in the persuasion threshold forces the sender to reveal more information to the receiver. In particular, when $\rho = 1$, the result always holds, namely raising q strictly increases equilibrium informativeness for both verifiability scenarios. This result allows us to consider an orthogonal test of commitment that, in contrast to Proposition 2, does not rely on changing ρ.

3.2 Experimental Design

Our laboratory implementation features many similarities with the one described in section 2. In particular, the monetary payoffs and the language used to describe the tasks are the same. Unlike section 2, however, the sender can now choose from among three messages, $M = \{r, b, n\}$. This additional richness of the message space allows us to easily switch between treatments with verifiable and unverifiable information, and makes these treatments more comparable. Two additional details of our implementation are worth mentioning. First, the revision stage is shown to the subjects only when it matters, namely, only for treatments with partial commitment, $\rho < 1$. For treatments with full commitment, instead, we avoid doing so to minimize confusion. Second, as in Section 2, we employ the strategy method at the guessing stage. That is, the receiver has to guess the color of the ball for *all* messages in the set M. In contrast, we do not use the strategy method for the revision stage. The sender revises only the part of her strategy that concerns the *realized* state . We do not elicit what the sender would have done had the ball been of a different color. In our view, this design choice achieves to goals: it makes the revision stage simple for our subjects and it highlights the stark contrast between revision and

commitment stage. This design choice, however, makes the computation of the correlation coefficients more challenging for treatments with partial commitment. We circumvent this problem of missing data by imputing the session-specific average behavior of the senders. This choice seems natural and, due to the random re-matching, receivers should hold comparable beliefs when facing a random sender in the last ten rounds of the experiment.[30]

In Appendix B and C, we present the instructions and provide examples of the graphical interface, which follows closely the one of Section 2.

3.2.1 Treatments and Equilibrium Predictions

In the experiment, we vary two main treatment parameters: the degree of the sender's commitment and whether or not information is verifiable. In treatments with verifiable information, the interface prevents senders from assigning positive probability to a red message conditional on a blue ball or to a blue message conditional on a red ball. The interfaces are identical in all other respects. For both verifiable and unverifiable information, we conduct three treatments with different degrees of commitment: $\rho \in \{0.20, 0.80, 1\}$. Thus, we have a total of six treatments forming a 2×3 factorial between-subjects design. We denote these treatments as illustrated by Table 2. Note that treatment $U100$ is nothing more than a variation on the benchmark treatment discussed in section 2, with the addition of the no message n. As can be seen in Table 3, this addition does not matter for the theoretical predictions.

Table 2: Treatments denominations

	Degree of Commitment		
Information	$\rho = 0.20$	$\rho = 0.80$	$\rho = 1.00$
Verifiable	V20	V80	$V100$
Unverifiable	U20	U80	U100

For each treatment, we conduct four sessions, for a total of 24 sessions. Each session included 12 to 24 subjects for a total of 384 subjects, who played 25 paid rounds (16 on average per session) in fixed roles. In addition to their earnings from the experiment, subjects received a $10 show-up fee. Average earnings, including

---

[30]Our results are robust to different imputation methods: For example, we can impute *subject*-specific averages and get essentially similar results. Also, it is important to note that the results for treatments with $\rho = 0.8$ (where we perform the imputation) are similar to those with $\rho = 1$ (where we do not need to use the imputation), suggesting the results are robust to our imputation method.

the show-up fee, were \$36.55, ranging from \$12 to \$60. On average, sessions lasted 100 minutes.

This experimental design allows us to capture many models of communication, ranging from cheap talk to disclosure and Bayesian persuasion. Note that, for two main reasons, we do not include the extreme cases with $\rho = 0$. First, these cases are the only ones for which experimental evidence already exists because these cases have received attention in the literature.[31] Second, the equilibrium predictions at $\rho = 0$ are identical to those at $\rho = 0.20$. Our main interest, instead, lies in treatments with partial or full commitment. These cases have never been tested in the lab and offer a unique opportunity to study the role of commitment in communication. Note, also, that our results for treatments with $\rho = 0.2$ are qualitatively consistent with prior observations from experiments with cheap talk and disclosure, that is, $\rho = 0$.

Table 3 reports the equilibrium predictions for each treatment in terms of the strategies played by senders and receivers.[32] Figure 7 reports the informativeness of equilibrium across our treatments.

We do not wish to go into the details of every case discussed in Table 3. However, we wish to emphasize that the equilibrium predictions displayed in Table 3 feature the key strategic tensions that we highlighted in Section 3.1.

First, the $V80$ and $U80$ treatments reveal a tension between the commitment and the revision stage, and this tension goes in opposite directions in verifiable versus unverifiable treatments. In treatment $U80$, anticipating their own behavior in the revision stage, senders are predicted to compensate by committing to reveal more information than in $U100$. By contrast, in treatment $V80$, senders are predicted to compensate by committing to reveal less information than in $V100$. Second, in both $V20$ and $U20$, the sender is unable to use commitment to undo her anticipated behavior at the revision stage. The predicted outcome in both these treatments is identical to the case of no commitment, $\rho = 0$. Therefore, we find a degree of indeterminacy: Not quite in the predicted level of informativeness, but rather in the actual strategies played by the subjects. Finally, as illustrated in Proposition 2, treatments $U100$ and $V100$ are predicted to induce the same outcomes (Figure 7). However, senders achieve this outcome by using substantially different strategies. Note also that the equilibrium of $U100$ is the same as in $U100S$; that is, the empty message plays no important role in the U treatment.

---

[31]See Blume et al. (2017) and the references therein.

[32]As discussed in section 3.1, the table presents the predictions assuming the specific equilibrium selection that we have made. Recall that, for the most part, multiplicity is about a selection of language, and all equilibria are equivalent in terms of payoffs and informativeness. The case with more substantive selection is V80. See section 3.1 for a discussion.
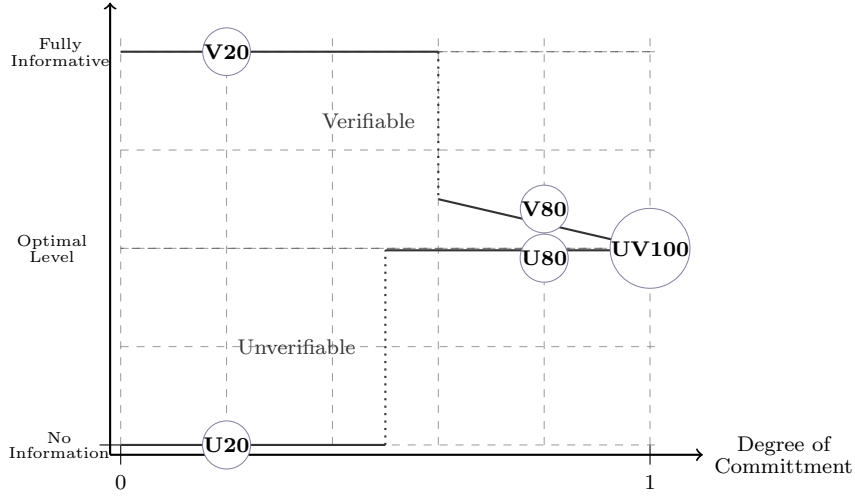
Figure 7: Predictions and Treatments.

4 Main Results

In this section, we present the main results of our experiment. Specifically, we present four sets of results.

First, we explore the simplest and most direct evidence to test whether subjects understand commitment and how they take advantage of it. To this purpose, we extensively exploit the flexibility of our experimental design. Our initial focus is on senders. We exploit the *within*-treatment variation between the commitment and the revision stage to track changes in their behavior. We then move to the study of receivers. In this case, we exploit the *across*-treatment variation and track how their responsiveness to information changes as we change the level of commitment ρ.

Second, we take on a more aggregate approach and we analyze how the amount of information that senders transmit changes, as we vary the level of commitment ρ. In doing so, we leverage a particular feature of our design. By Proposition 2, the predicted changes in informativeness as a function of ρ have opposite signs depending on whether information is verifiable. These contrasting comparative statics allow a particularly tight test of the role of commitment in communication.

Third, we zoom-in on a pair of treatments that are of particular interest, namely, $V100$ and $U100$. As explained in Proposition 2, these treatments are somewhat special because the equilibrium outcome is rule-independent. Yet the strategies leading to these identical outcomes can be radically different because of the role played by the two different rules we consider. This environment is a particularly natural one in which to learn about the way rules shape subjects' incentives and behavior in the laboratory. The discrepancy between observed behavior and theoretical prediction could be relevant for policy. Fourth, we use the predictions from Proposition 3 to

Table 3: Equilibrium Predictions

		Sender							Receiver		Correlation
		Commitment				*Revision*			*Guessing*		Coefficient
Treat.	Ball	Message			Ball	Message			Mes.	Guess	ϕ
		red	blue	no		red	blue	no			
V20	R	1		0	R	1		0	red	*red*	
	B		x	$1-x$	B		x	$1-x$	blue	*blue*	1
									no	*blue*	
V80	R	0		1	R	1		0	red	*red*	
	B		$\frac{3}{4}$	$\frac{1}{4}$	B		0	1	blue	*blue*	0.57
									no	*red*	
V100	R	0		1					red	*red*	
	B		$\frac{1}{2}$	$\frac{1}{2}$					blue	*blue*	0.50
									no	*red*	
U20	R	x	y	$1-x-y$	R	1	0	0	red	*blue*	
	B	x	y	$1-x-y$	B	1	0	0	blue	*blue*	0
									no	*blue*	
U80	R	1	0	0	R	1	0	0	red	*red*	
	B	$\frac{3}{8}$	$\frac{5}{8}$	0	B	1	0	0	blue	*blue*	0.50
									no	*blue*	
U100	R	1	0	0					red	*red*	
	B	$\frac{1}{2}$	$\frac{1}{2}$	0					blue	*blue*	0.50
									no	*blue*	

x and y indicate any (feasible) probability.

understand how senders react to changes in the persuasion threshold q.

4.1 Response to Commitment

The assumption of commitment is a defining feature of persuasion models and represents the main departure relative to cheap-talk and disclosure models. We now discuss the degree to which subjects react to the availability of commitment.

4.1.1 Senders and Commitment

We begin by focusing on senders' behavior. We first exploit within-treatment variation to evaluate the role of commitment in shaping senders' behavior. We do so by comparing behavior in the commitment stage with behavior in the revision stage, and exploiting the fact that this comparison changes depending on the verifiability of information. For example, when $\rho = 0.8$, the predicted behavior in the commitment stage displays particularly stark differences relative to the behavior in the revision stage (see Table 3). This within-treatment variation provides us with a very simple test that we use to evaluate the extent to which senders understand the role of commitment in this game, and whether they are able to use it to their advantage.

In Figure 8, we present the average difference in senders' strategies between the
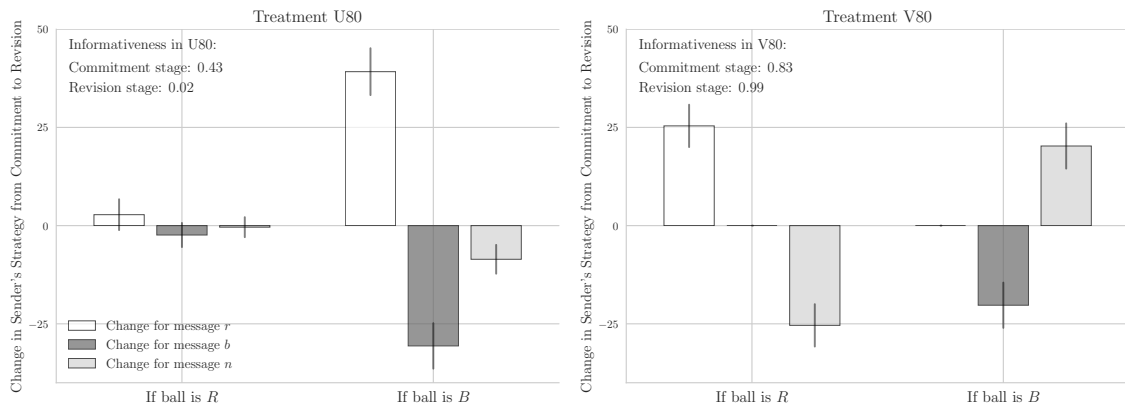
Figure 8: Sender's Strategy: Commitment vs. Revision, $\rho = 0.8$

commitment stage and the revision stage for $U80$ and $V80$. A high bar indicates a message that is sent more often in the revision stage, a *negative* bar a message that is used more in the commitment stage. This exercise is done separately by state. Let us consider the first three bars (from the left), that is treatment $U80$ when the ball is R. Using Table 3, we can see that the strategy in the revision and commitment stages are the same, hence all three bars should be of zero height. This outcome is exactly what we observe in the figure. However, when the ball is B, Table 3 tells us that the frequency of r should be higher in the revisions stage (1 versus $\frac{3}{8}$), while the frequency of b should increase, and the n message should not change. In the data we do observe the predicted increase in r and decrease in b. In contrast with the theory, there is a small change in n, but this difference is insignificant. Turning to $V80$ when the ball is R, Table 3 indicates that the sender should replace r messages from the commitment stage with n messages in the revision stage. This is exactly what we see in the right panel of Figure 8. Finally, when the state is B, Table 3 predicts that n messages should increase and b messages decrease. Once again, this is exactly what appears in the last three bars (to the right) of Figure 8. The changes in frequencies are jointly significant in each panel and for each state ($p < 0.01$ in both cases).[33]

An alternative way to confirm that the observed changes in behavior are in line with the theoretical prediction is to compare the informativeness contained in the commitment strategy with that contained in the revision strategy. Denote these two quantities by ϕ_C^B and ϕ_R^B, respectively. In line with the prediction of Proposition 1, in $U80$ senders do reveal substantially more information in the commitment stage, $\phi_C^B = 0.43$, than in the revision stage, $\phi_R^B = 0.02$. This difference is significant ($p < 0.01$) and quantitatively large. Also consistent with the theoretical prediction, we observe that in $V80$ informativeness in the commitment stage, $\phi_C^B = 0.83$, is *lower* than that of the revision stage, $\phi_R^B = 0.99$ (statistical significance at $p < 0.01$).

---

[33]Although statistically significant changes occur for $U80$ when the state is R, they are small in magnitude.

Summing up, the joint evidence coming from treatments $U80$ and $V80$ suggests that senders react to commitment and do so in ways that are consistent with the theory. Our evidence suggests that many senders exploit their commitment power to strategically hide good news (state R) when information is verifiable, and hide bad news (state B) when information is unverifiable. From a quantitative point of view, these efforts by the senders fall short of exactly matching the equilibrium predictions. Most of the deviations from equilibrium come from behavior in the commitment stage. In contrast, behavior in the revision stage is quite close to the theory. We discuss departures from equilibrium in more detail in section 4.3. Qualitatively, however, this central prediction of our strategic communication model is corroborated by the data. We emphasize that the predictions are completely different for verifiable and unverifiable messages. For instance the prediction on the change in informativeness between the commitment and the revision stage go in opposite directions. This is a useful feature of considering verifiable and unverifiable messages in a similar framework, as they allow us to consider very different predictions within the same environment. The fact that these predictions are matched in the data should be all the more reassuring about the ability of senders to use commitment as predicted by the theory.

4.1.2 Receivers and Commitment

We now focus on receivers: Our goal is to evaluate the extent to which they understand the strategic implications of commitment, and whether their reactions are consistent with the theory. To do so, we create a direct test that is specifically tailored to the problem they face. Consider the Bayesian posterior conditional on a message m, computed only on the basis of the information contained in the commitment strategy π_C.[34] This posterior belief, which we will call *interim posterior*, can be interpreted as the belief a receiver would hold if she ignored the existence of a revision stage. Clearly, when $\rho = 1$, interim and ex-post beliefs coincide. More generally, given π_C and π_R, the higher the degree of commitment ρ, the closer the interim posterior is to the ex-post one that conditions on all the information, including the sender's equilibrium behavior in the revision stage. We use this simple observation to test whether receivers understand the strategic implications of different levels of commitment. Thus, we should observe *different* guessing behavior at *identical* interim beliefs for *different* degrees of commitment. In particular, at high levels of commitment, interim beliefs should be highly influential in guiding receivers' behavior; at low levels of commitment, they should not.

This analysis is carried out in Figure 9. We begin by comparing treatments $U20$ and $U100$ (left panel). We focus on the interim posteriors after message r, which is the key *strategic message* in this treatment. We compare how receivers respond to

---

[34]That is, $\mu_0(R)\pi_C(m|R)/(\sum_\theta \mu_0(\theta)\pi_C(m|\theta))$.

this message as a function of the induced interim posterior and of the treatment.[35] In $U20$, the interim posterior should have little or no impact on the receiver's guess, because the message most likely did not come from the commitment strategy, and therefore the interim posterior is likely to be far from the final posterior. By contrast, in $U100$, the interim posterior should have a substantial effect on the receiver's guess. In particular, the receiver should guess *red* for high-enough posteriors. In fact, because the message came from the commitment strategy with probability one, the interim belief coincides with the ex-post belief. Consistently with these predictions, the estimated receivers' response in the left panel of Figure 9 is mostly flat in $U20$, whereas it is strictly increasing in $U100$.

Similar, if not stronger, evidence is found when comparing $V20$ and $V100$. By the nature of verifiable information, messages r and b induce trivial interim beliefs of either 1 or 0, respectively. The message that potentially entails strategic considerations is message n, and this message is the one we focus on. As can be seen in the right panel of Figure 9, the estimated receivers' response to an increase in the interim posterior is weak in treatment $V20$, whereas it is strong and positive for $V100$.[36] Overall, the joint evidence coming from treatments with verifiable and with unverifiable information suggests that, on average, receivers understand the basic strategic implications of the role that commitment plays in our model and react to it in ways that are broadly consistent with the theory.

Another striking feature of Figure 9 emerges from the comparison between receivers behavior in U100 versus V100. The response to message r in U100 is almost identical to the response to message n in V100. Thus, receivers react to the "persuasive" message in the same way in the two treatments, despite the fact that the nature of the message is quite different in the two treatments: it is consistent with natural language in U100, whereas it is not in V100.

4.2   Cross-Treatment Comparisons

The previous sub-section establishes that both senders and receivers react to the power of commitment. This section explores additional dimensions of the degree to which agents react to commitment in ways that are consistent with the theory. One key prediction of the theory, stated in Proposition 2, concerns how equilibrium

---

[35]In Figure 9, the solid lines are the polynomial fit of the induced interim posterior and the observed guess.

[36]The probability that the receiver guesses *red* when the interim posterior is below $\frac{1}{2}$ does not differ statistically between $\rho = 0.2$ and $\rho = 1$, both for the case with unverifiable information (left panel) and verifiable information (right panel). Note that for the case of verifiable information, the difference can be significant depending on how the test is performed. For interim posteriors above $\frac{1}{2}$, we find a statistically significant difference in both cases ($p < 0.01$ in both cases) and, perhaps more importantly, the magnitude of the change is much more sizable: 56 versus 14 percentage points in the verifiable case, and 40 versus 6 percentage points in the unverifiable case.
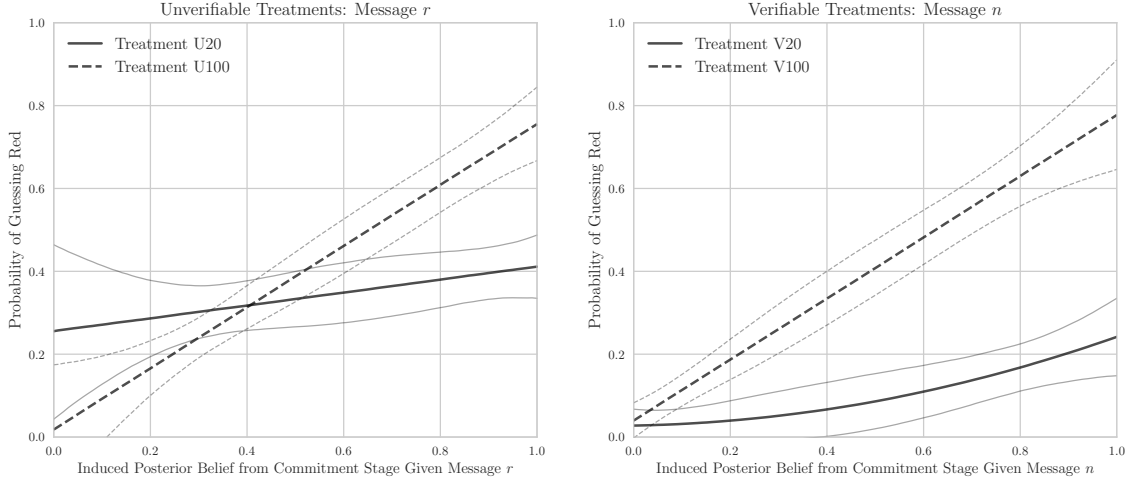
Figure 9: Receiver's Response to Persuasive Messages: $\rho = 0.2$ vs. $\rho = 1$

informativeness changes with commitment under verifiable and unverifiable information. Figure 10 shows the distributions of sender-specific informativeness ϕ^B in our main treatments. We wish to highlight two patterns that emerge from this figure. First, we find a noticeable first-order stochastic *increase* in the distribution of informativeness in $U100$ relative to $U20$ (left panel), as well as in $U80$ relative to $U20$. Thus, under unverifiable information, the amount of information transmitted by the senders increases as commitment increases, as predicted by the theory. Second, we see a first-order stochastic *decrease* in the distribution of informativeness in $V100$ relative to $V20$ (right panel), and less so in $V80$ relative to $V20$. Thus, under verifiable information, the amount of information transmitted by the senders decreases as commitment increases, as predicted by the theory.[37]

Although these patterns are qualitatively in line with theory, we note sizable quantitative deviations from the point-predictions of the theory. To illustrate this finding, it is sufficient to compare the average informativeness by treatment. For each treatment, Table 4 reports the theoretical informativeness $\phi^\star$ in the top left panel, empirical informativeness ϕ in the top right panel, and informativeness with hypothetical Bayesian receivers ϕ^B in the bottom left panel. The differences between ϕ^B and ϕ allow us to partly disentangle whether senders or receivers are primarily responsible for possible deviations from the equilibrium. The symbols between the numbers indicate their "statistical relations" at the 10% level; that is, $\approx$ means that the p-value of the equality of the two is larger than 0.1.

In the top-right panel of the table, we see that subjects react to commitment in the expected direction, both for verifiable and unverifiable information. However,

---

[37]As predicted by the theory, $U80$ and $U100$ are unranked. The same is true, although to a lesser extent, for the comparison between $V80$ and $V100$. Finally, we note that the CDF for $U100S$ is similar to that for $U100$, as predicted by the theory. The two are plotted together in the left panel of Figure D28 in the appendix.
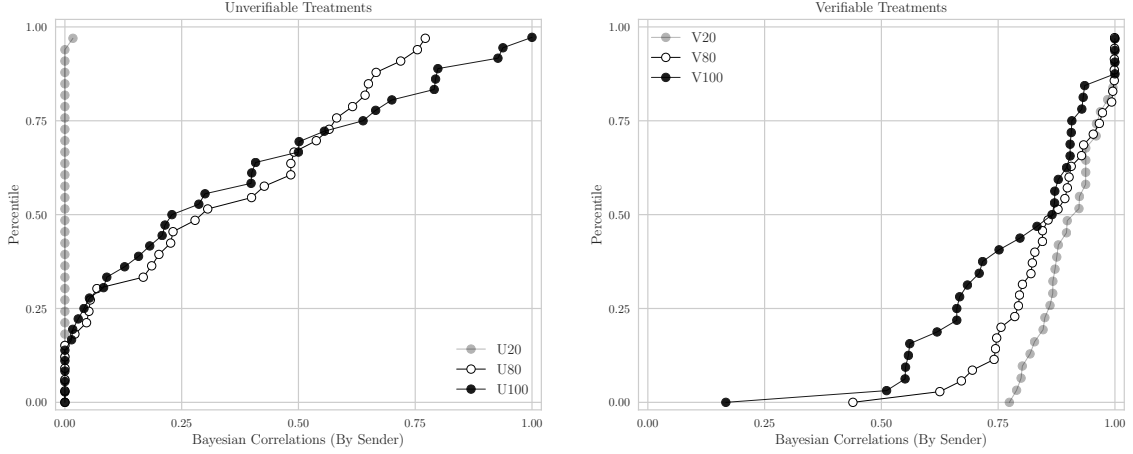
Figure 10: CDF of Subject Average Bayes Correlation ($\bar{\phi}_i^B$) by Treatment

Table 4: Average Correlations per Treatment

$\phi^\star$ – Theoretical Predictions

	Commitment (ρ)		
	20%	**80%**	**100%**
Verifiable	1	0.57	0.50
Unverifiable	0	0.50	0.50

ϕ – Empirical Correlation

	Commitment (ρ)				
	20%		**80%**		**100%**
Verifiable	0.83	$\approx$	0.78	$>$	0.68
	$\vee$		$\vee$		$\vee$
Unverifiable	0.09	$<$	0.20	$\approx$	0.22

ϕ^B – Empirical Correlation with Bayesian Receivers

	Commitment (ρ)				
	20%		**80%**		**100%**
Verifiable	0.89	$\approx$	0.85	$>$	0.78
	$\vee$		$\vee$		$\vee$
Unverifiable	0.00	$<$	0.33	$\approx$	0.34

Note: black symbol, as predicted;
gray symbol, not as predicted.

the observed changes are more muted than what is predicted by the theory. In the case of verifiable information, for example, the theory predicts that, moving from $V20$ to $V100$, we should observe a reduction of 0.5 in the correlation (from 1 to 0.5). In the data, the corresponding reduction is 0.13, or only 26% of the predicted change. Similarly, under unverifiable information, the changes are in the predicted directions, although the magnitudes are smaller. Comparing these correlations with those on the bottom-left panel suggests only part of the "missing effect" can be imputed to "noise" introduced by non-Bayesian receivers. Recall that when we replace our actual receivers with a hypothetical Bayesian receiver, we effectively shut down the dampening effect that receivers' mistakes produce on the correlation. A receiver's behavior that becomes noisier can, in fact, only reduce the correlation ϕ, not increase it. The changes in ϕ^B reveal larger effects of commitment for the unverifiable treatments and smaller effects for the verifiable treatments: 68% in the case of the unverifiable treatments and 22% of the predicted change for verifiable treat-
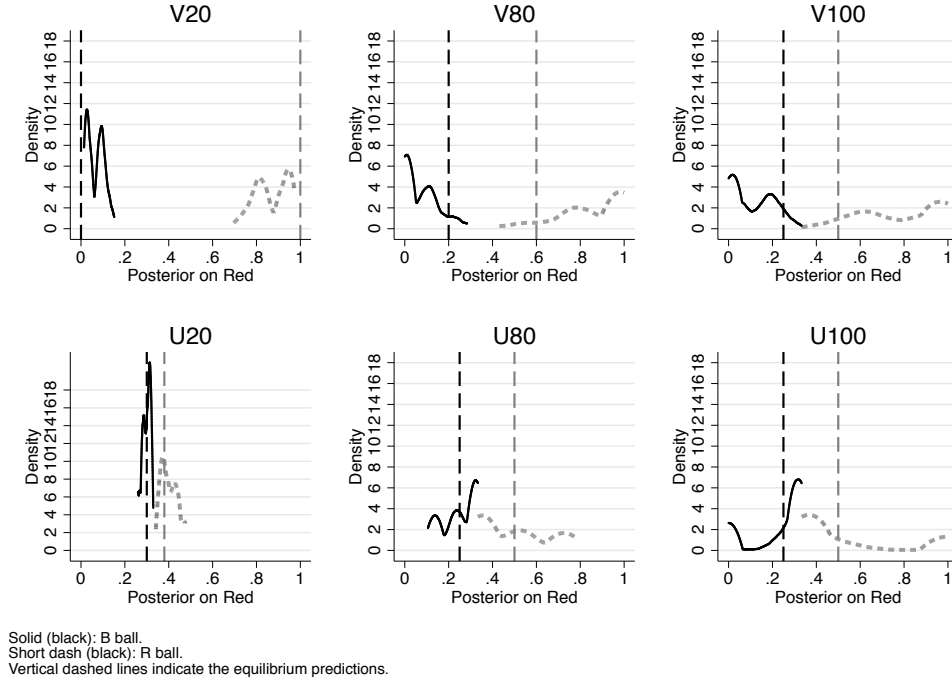
Solid (black): B ball.
Short dash (black): R ball.
Vertical dashed lines indicate the equilibrium predictions.

Figure 11: Posterior on R as a Function of the State

ments. However, especially for the unverifiable treatments, many senders clearly generate correlations that are positive, but too low to be persuasive, that is, to induce a receiver to choose *red*.

To understand this phenomenon better, we now turn to an analysis of the posteriors that senders induce with their communication strategies. In particular, Figure 11 displays the kernel density estimates of the Bayesian posteriors *conditional on the state*.[38] The vertical dashed lines indicate the theoretical predictions; the other lines present the data under the different treatments.[39] For instance, for the treatments with $\rho = 1$, the vertical long-dash gray line is at 0.5 because in equilibrium, the posterior following the red state is 0.5. The vertical long-dash black line is at 0.25 because in equilibrium, in the blue state, the posterior is 0.5 with 50% probability (when the sender sends the r message) and 0 with 50% probability (when the sender sends the b message).

In all cases we see a sizable response to the treatment in the direction predicted by the theory, more so in the unverifiable than in the verifiable treatments. Moving from U20 to U100, the posteriors become more spread out, whereas moving from V20 to V100, the posteriors move closer, as predicted by theory. However, there are important discrepancies: whereas in the case of V20, the posteriors are inside the

---

[38]Given the state and strategy in both the commitment and the revisions stage, we compute the expected posterior conditional on the likelihood of each message.

[39]In the bottom left panel, the two vertical lines are not both at 1/3 because they are computed assuming senders reveal all the information in the commitment stage.

Table 5: Difference In Mean State Conditional Posteriors
(theoretical values in parentheses)

	Commitment (ρ)					
	20%		**80%**		**100%**	
Verifiable						
Difference:	0.80		0.78		0.69	
	(1.00)		(0.40)		(0.25)	
	B	**R**	**B**	**R**	**B**	**R**
Mean:	0.07	0.87	0.07	0.86	0.10	0.79
Unverifiable						
Difference:	0.11		0.24		0.30	
	(0.00)		(0.25)		(0.25)	
	B	**R**	**B**	**R**	**B**	**R**
Mean:	0.30	0.41	0.25	0.49	0.23	0.53

lines describing the theoretical predictions, in the other two verifiability treatments, most of the mass of the posteriors lies outside of the relevant (theory-predicted) lines. In other words, senders are not informative enough under V20, and are too informative in the other cases.

Table 5 reports the difference between the mean posteriors when the ball is red relative to the case in which the ball is blue. The table shows that the data move in the right direction for both verifiable and unverifiable treatments, but that the mean difference is much closer to the theoretical predictions in the case of the unverifiable treatments than in the case of verifiable treatments.

According to the data presented in Table 5, the posterior difference is very close to predicted in treatments U80 and U100 but quite far in treatment V100. Recall that in theory, the treatments U100 and V100 should yield the same equilibrium outcomes.

We now summarize. The behavior of our subjects is not, on average, in line with the point predictions of the theory. However, our data qualitatively match the asymmetric theoretical predictions that our framework produces: Increasing commitment has opposite effects on information transmission under verifiable and unverifiable information. Interestingly, rules have a greater impact than predicted when commitment is high.

4.3 The Impact of Rules

Regarding the effect of verifiability, our model makes two predictions. First, for each given level of commitment, the amount of information conveyed by senders is

(weakly) higher under verifiable information than unverifiable information. Second, under full commitment the amount of information (and the equilibrium outcome) is independent of verifiability: with full commitment, *rules* should not matter (Proposition 2).

The first of these predictions is strongly borne out in the data: for every given level of commitment, more information is conveyed by the senders under verifiable treatments than under unverifiable treatments. On the sender side, Table 4 reports that the correlation ϕ^B is 0.89 in V20 and 0.00 in U20 (predictions are 1 and 0, respectively). On the receiver side, the probability of guessing *red* following message r (resp. b) across all treatments with verifiable information is 95% (resp. 1%), thereby suggesting that receivers correctly understand the implications of verifiable information.[40]

However, the second prediction is strongly rejected by the data. There is a sizable observed difference in the level of informativeness between treatments U100 and V100. The average Bayesian correlation is 0.78 in V100 and 0.34 in U100 (Table 4). More broadly, as reported in Figure 12, the discrepancy between the two treatments is substantial, affecting all quantiles of the distribution of informativeness ρ^B.[41]
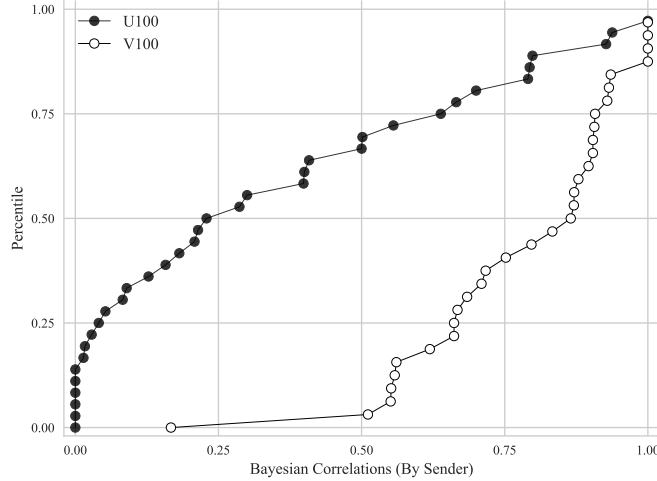

Figure 12: CDF of Subject Average Bayesian correlation ($\bar{\phi}_i^B$): V100 and U100

In order to explain this discrepancy between the theory and the data, we look more closely into the behavior of senders and receivers in these two treatments. It is apparent that the biggest difference between V100 and U100 lies in the much more prevalent use of uninformative strategies in U100. Figure 12 illustrates that

---

[40]The comparable numbers in treatments with unverifiable information are 40% and 8%.

[41]Furthermore, even for cases of partial commitment the difference in informativeness in our data is quantitatively larger than predicted by the theory.

no sender in V100 uses such strategies. In contrast, 16% of senders in U100 are consistently uninformative. A similar conclusion holds if we use different cutoffs for ϕ_i^B. Overall, uninformative communication strategies are uncommon in V100, whereas they are quite common in U100. Note that fully uninformative strategies are just as feasible in V100 as in U100: a sender can produce no information to the receiver by sending message n in both states.

Figure 13 presents a clustering analysis of the senders' strategies for treatment V100 that is analogous to the one we presented in Figure 6 for U100S in Section 2.2.[42] There are three main clusters that emerge from this analysis. First, there is a large cluster that is suggestive of equilibrium behavior. Senders use message n with high probability when the state is R and randomize between b and n when the state is B. Second, there is cluster that is fully informative in the natural language. Senders are truthful and send message r when the state is R and b when the state is B. Third, there is a cluster that, like the previous one, is also fully informative, but senders use message n instead of b, when the state is B. When comparing senders' behavior in V100 and U100, we note that the first two types of clusters, "equilibrium-like" and "truth-telling," appear in both treatments. The biggest qualitative difference is in the third cluster. When information is unverifiable, a cluster of senders is uninformative, and sends message r in both states. As we saw above, this behavior is absent in V100. It is replaced by a cluster of fully informative and yet non-truthful senders.
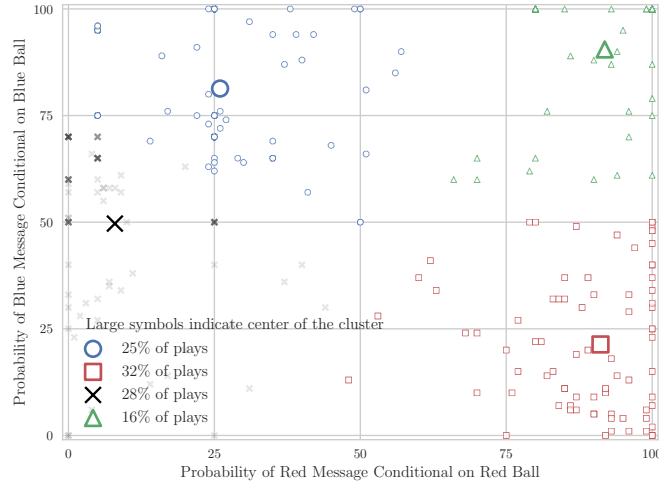

Figure 13: Sender's Strategies in V100 Grouped in Clusters

One potential explanation for this difference in behavior is driven by a specific misunderstanding by some of the senders: namely, these senders act as if they were secretly choosing the messages instead of publicly choosing the information struc- tures. In other words, they act as if they were choosing strategies for a hypothetical

---

[42]The clusters for U100 are quite similar to those for U100S. In order to save space, we do not repeat this analysis here.

revision stage. Observed receivers' behavior yields very different incentives in V100 versus U100 under this behavioral assumption about these senders. The optimal strategies are exactly the ones described above. In the unverifiable case r is optimal in both states; in the verifiable case r is optimal in the R state and n is optimal when the state is B (because n leads to much higher probability of the receivers being persuaded).[43]

More generally, senders' behavior is much more informative in V100 than U100. We discuss two additonal possible explanations that relate to how senders react to receivers behavior.

A first hypothesis related to receivers behavior is motivated by the literature on the disclosure of verifiable information without commitment. Perhaps receivers are overly skeptical of message n in V100, leading senders to switch and use the verifiable message r. After all, this skepticism is they key force pushing toward disclosure in standard games of disclosure with verifiable messages without commitment.[44] However, in our environment senders have full commitment, and this makes all the difference. Indeed, as we discussed in Section 4.1.2, the data highlighted in Figure 9 shows that receivers respond to message r in U100 and to message n in V100 in extremely similar ways, even when controlling for implied Bayesian beliefs. Therefore, the discrepancy in senders' behavior cannot be rationalized by this particular difference in receivers' behavior.

A second hypothesis that focuses on receivers' behavior is that, perhaps, full disclosure is more profitable in V100 than in U100, despite full commitment. This could be the case if disclosure in V100 is trusted more by receivers, due to information verifiability.[45] This hypothesis does receive some support in the data: senders' earnings from full disclosure are indeed higher in V100 than in U100.

The discrepancy in senders' behavior documented in this section is potentially important for policy because it suggests a novel role for information verifiability. In contrast with the literature on the failure of the unraveling principle, we find that under information verifiability senders transmit an excessive amount of information compared to the predictions of the theory. Despite the presence of commitment power, we find that many senders are unable to strategically withhold information to their advantage. Moreover, the large amount of uninformative behavior by the sender that we observe in U100 is mostly associated with sending r rather than n.

---

[43]This behavior is in fact what we observe in the data in the revision stages of U80 and V80 respectively. In the revision stage of V80, the median probability of choosing r in state R is 0.99, and the median probability of choosing n in state B is 0.74. In the revision stage of U80, the median probability of choosing r in state R is 0.91; and the median probability of choosing r in state B 0.82).

[44]Experimentally, it is indeed found that no news is bad news for receivers, although receivers are not sufficiently skeptical. See for instance Jin et al. (2016).

[45]This is particularly evident from Figure 2 in Section 2.2. Even when provided with conclusive evidence that the state is R, some receivers in U100S still do not trust the sender.

This behavior suggests that this finding is unrelated to senders' lying aversion.

4.4 Changing the Receiver's Incentives

One of the more direct implication of commitment, which is stated in Proposition 3, is that increasing q leads to more informative communication from senders. Based on this idea, we designed one more treatment, in which it is more valuable for the receiver to match the state when the state is B. This treatment provides a different test of whether subjects react to the power of commitment. This treatment involves full commitment ($\rho = 1$) and unverifiable information and is referred to as treatment $U100H$. In this treatment, we only change payoffs so that receivers require more persuasion in order to choose the senders' favorite action. Payoffs are as follows. As in all other treatments, the receiver obtains zero payoff if he makes the wrong guess. In contrast to the treatments above, the receiver wins different amounts if he correctly guesses the color of the ball: 2 if the ball is Blue, $\frac{2}{3}$ if ball is Red. The sender wins 3 if the receiver guesses Red. With this new treatment, the persuasion threshold (the posterior above which the receiver guesses red) changes from 0.5 to 0.75. Thus, in equilibrium, the sender should provide more information. The strategy of the sender involves sending r with probability one if the ball is Red, sending r with probability 1/6 and b with probability 5/6 if the ball is blue.

For the U100H treatment, we conducted four sessions, each with 16-20 subjects (72 in total). Those subjects made between \$10.48 and \$26.56 (average \$18.02).
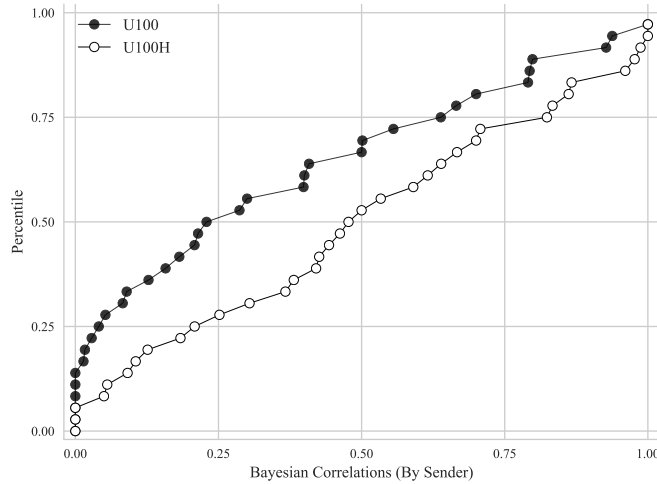

Figure 14: CDF of Subject Average Bayesian Correlation ($\bar{\phi}_i^B$): U100 and U100H

Figure 14 shows the distribution of ϕ^B for the U100H treatment is to the right of that for the U100 treatment. The median subject goes from inducing an average

correlation of 0.22 in U100 to one of 0.47 in U100H. This shift, however, is not statistically significant ($p > 0.1$). We note, though, that sender behavior in the U100H treatment evolves considerably over time, as can be seen in Figure D19, and if we regress the Bayesian correlation on a dummy for the U100H treatment, but also add match variables interracted with treatment dummies; the match-interaction variable is significant and positive for the U100H treatment ($p < 0.01$).[46] Hence, a wedge is indeed building over time, with senders ultimately conveying more information in the U100H treatment than in the U100 treatment.

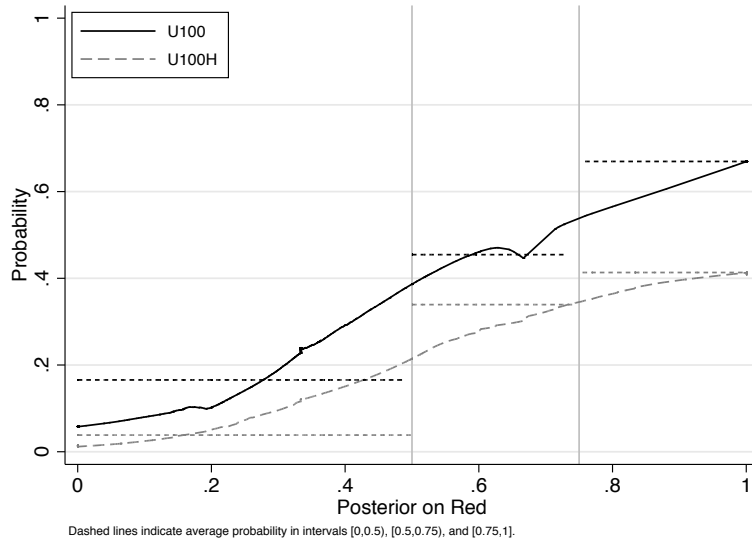Dashed lines indicate average probability in intervals [0,0.5), [0.5,0.75), and [0.75,1].

Figure 15: Probability of Guessing *red* by Posterior for treatments U100 and U100H

The theory suggests senders need to send more information to convince receivers to listen to them. Figure 15 shows our receivers also require more evidence in the U100H treatment than in the U100 treatment (smoothed line and averages for key intervals). As predicted, receivers are more likely to guess *red* for posteriors between 0.5 and 0.75 ($p < 0.05$) in U100H. However, this effect is also found for posteriors below 0.5, which is not predicted ($p < 0.1$). But, as predicted, the average difference for posteriors at or above 0.75 is not significant ($p > 0.1$).[47]

4.5 Quantal Response Equilibrium

Prior experimental work on communication games has explored, among others, two approaches to explain departures from equilibrium play: Level-k and quantal response equilibrium (QRE).[48] In this section we explore a QRE analysis of our treat-

---

[46]To account for time, this test does not average at the subject level.

[47]Appendix D provides additional information on the behavior of receivers in Figure D27.

[48]For Level-k, see, for instance, Nagel (1995). For QRE, see Goeree et al. (2016). In the experimental literature on cheap-talk games, Cai and Wang (2006), Kawagoe and Takizawa (2009)

ments with full commitment and discuss why we adopt this approach rather than a Level-k one.[49]

Unfortunately, a straightforward analysis of Level-k is not helpful in our setting. A key component of a Level-k analysis is the specification of level-0 players. First, in any of the full-commitment treatments, with or without verifiable information, the strategy of the sender is fully observable by the receiver, so there is little room for lack of strategic sophistication on the part of receivers to play a role. Second, in all our verifiable treatments, there is no leeway in specifying receivers' beliefs (or behavior) following a red message or a blue message. Thus, the only degree of freedom is in specifying non-equilibrium beliefs and play following a "no message." The natural assumption is that a level-0 receiver responds to such messages in a naive way, by failing to update her beliefs and assigning a posterior of 1/3, the same as the prior. However, in our setting, such belief yields the same optimal response (blue) as the equilibrium beliefs following a b message. The fact that receivers' behavior is identical between level-0 and equilibrium play implies that this concept, taken as is, gives us little leverage to explain departures from equilibrium in our environment. This is of course not to say that a more elaborate version of Level-k, possibly combined with other approaches, may not be a fruitful avenue to explore. However, we chose not to pursue such elaborate alternative approaches. We instead developed in detail an analysis of QRE for treatments with full commitment.

The approach taken in QRE is to assume that players respond with errors to their beliefs and that, in equilibrium, these beliefs correctly account for the errors that other players make. Despite the simplicity of our design, estimating QRE in a multi-stage game with incomplete information and a continuum of actions is particularly challenging. To address these challenges we adapt the standard QRE methodology to our problem in the following way. First, let Π_k be a given set containing k different sender's strategies, i.e. information structures $\pi : \Theta \to \Delta(M^\theta)$. Later, we explain how this set is determined in our procedure. For each $\pi \in \Pi_k$ and $m \in M^\theta$, let $\mu_\pi(\theta_H|m)$ be the posterior belief on θ_H, given m and π.[50] Denote by $U(a_H|\pi, m) = \mu_\pi(\theta_H|m)$ (resp. $U(a_L|\pi, m) = 1 - \mu_\pi(a_L|m)$) the expected utility of choosing a_H (resp. a_L). The Logit QRE model assumes that a receiver of type λ_R

and Wang et al. (2010) have explored Level-k or QRE models to explain deviations from equilibrium predictions.

[49]There are two reasons why we focus on the cases with full commitment. First, these are the cases that display the most striking departures from equilibrium predictions. Second, the estimation with partial commitment presents substantial additional challenges, not the least of which being the sizable increase in the strategy space for the senders.

[50]Receiver behavior conditional on messages that have zero probability is irrelevant for the estimation of QRE parameters (λ_S, λ_R). Therefore, in this section, we can ignore the fact that $\mu_\pi(\theta_H|m)$ may be not well-defined at all histories.

chooses action a_H with the following probability:

$$\mathbb{P}_R(a_H|\pi, m, \lambda_R) = \frac{e^{\lambda_R U(a_H|\pi,m)}}{e^{\lambda_R U(a_H|\pi,m)} + e^{\lambda_R U(a_L|\pi,m)}}.$$

Given λ_R, the sender's expected utility from choosing π is given by vector $V(\pi|\lambda_R) := \sum_{\theta,m} \mu_0(\theta)\pi(m|\theta)\mathbb{P}_R(a_H|\pi, m, \lambda_R)$. That is, the sender takes receivers' errors into account when computing her expected payoff from playing a certain strategy. The probability that a sender of type λ_S chooses $\pi \in \Pi_k$ is given by

$$\mathbb{P}_S(\pi|\lambda_S, \lambda_R) = \frac{e^{\lambda_S V(\pi|\lambda_R)}}{\sum_{\pi \in \Pi_k} e^{\lambda_S V(\pi|\lambda_R)}}.$$

Note that, in treatments with full commitment, the receiver perfectly observes the strategy π chosen by the sender. Whether or not this strategy was chosen by mistake is irrelevant for the receiver who, instead, best responds to π and m as described above. Effectively, the receiver solves a single-agent decision problem. Therefore, we can estimate $\hat{\lambda}_R$ from the data using MLE, irrespective of λ_S. The sender, instead, moves before the receiver and therefore must take into account λ_R. For each strategy $\pi \in \Pi_k$, we use the senders' *empirical* expected payoffs to consistently estimate $V(\pi|\lambda_R)$ (see Bajari and Hortacsu (2005)). Given these values, we use MLE to estimate $\hat{\lambda}_S$. The parameter λ_i captures how well a player best-responds to her beliefs about her opponent's behavior. At one extreme, as $\lambda_i \to \infty$, players become perfectly rational. At the other extreme, when $\lambda_i = 0$, players randomize uniformly across available actions.[51]

This estimation procedure clearly depends on the initial choice of Π_k, a finite set of strategies for the sender. Because the strategy space is multi-dimensional and senders' strategies are not evenly distributed on this space, creating a partition of the data is complex. To determine Π_k, we use a k-means clustering algorithm–the same algorithm we used in Section 2.2. This algorithm computes exactly k clusters, by minimizing the distance between each strategy that belongs to a cluster and its mean. In what follows, we highlight the results that become robust as the number of clusters becomes large enough. When the number of clusters is too small, choices with very different expected payoffs are pooled together, which can lead to results that are not meaningful. The estimates that we report in Table 6 are computed for $k = 22$.[52]

A particularly useful feature of our setting (and our procedure for estimating QRE) is that it allows us to make meaningful comparisons across treatments by

---

[51]Note that we let $\lambda_S \neq \lambda_R$. One could impose an additional restriction requiring that $\lambda_R = \lambda_S$ and estimate these two parameters simultaneously. We do not impose this restriction as it would contrast with the fact that senders and receivers face drastically different tasks in our game.

[52]The qualitative results highlighted below are true for all of our estimates with 8 or more clusters.

Table 6: QRE λ Estimates

Treatment	Sender	Receiver
U100	0.36	1.31
V100	1.99	1.79
U100S	2.11	1.54
U100H	2.34	1.23

comparing the estimated λs. In contrast, in most experiments, the risk preferences of the subjects are unknown and the task that subjects face can vary substantially across treatments. Thus, comparisons of QRE estimates across treatments is sometimes difficult to interpret. In our design, instead, because outcomes are binary, risk preferences are irrelevant. Treatments with full commitment only differ in the restrictions imposed on the senders' message spaces. However, the relevant space for determining outcomes is the space of induced posteriors, and this space can be partitioned in the same way for all treatments with full commitment.[53]

Comparing U100 and V100 reveals that both senders and receivers are closer to best responding to one another's behavior in the V100 treatment than in the U100 treatment. This suggests that rules simplify the problem or make incentives more transparent for the subjects. Similarly, comparing U100 to U100S reveal estimates that are higher for both roles in the U100S treatment (although only slightly for the receiver). This result is reasonable given that U100S is a simplified version of U100. When comparing U100 and U100H, we have that the sender's estimate is higher in U100H but slightly lower for the receiver. The higher estimate for the sender in U100H could be explained by the fact that the incentives to convey information are "steeper" in that treatment, and indeed there is less babbling in that case.

5   Conclusion

This paper explores whether experimental subjects recognize, and react to, the power of commitment in communication. To this end, we use the fact that commitment has opposite effects on information transmission when messages are verifiable versus the case in which they are not. Indeed, when messages are unverifiable, increasing commitment allows senders to convince receivers of the credibility of their messages and to improve upon the babbling equilibrium of cheap-talk games. However, when

---

[53]In Table 6, we report QRE estimates that are computed by letting Π_k vary across treatment. This is consistent with what we have presented up to this point and keeps the methodology transparent and simple to understand. However, we show in the Appendix that the qualitative results we highlight here are robust to clustering the senders' strategy space in the same way for all treatments. Furthermore, the qualitative results in that case are true for all estimates with 8 or more clusters. The complete results for different number of clusters and for the case where the clustering is performed in the same way are presented in the appendix.

messages are verifiable, increasing commitment allows senders to undo the unravelling that happens in a standard disclosure environment, and to withhold some of the information. When commitment is partial but high enough, our implementation allows us to directly observe whether senders recognize the role commitment. We can also study whether receivers recognize the implications of commitment for the content of messages. In addition, we explore the reaction to changing the persuasion threshold, which is one more way to examine whether subjects react to commitment as predicted. Finally our experiment provides one of the first experimental investigations of Kamenica and Gentzkow (2011).[54]

Our findings suggest that the central force at play in Kamenica and Gentzkow (2011) is one that many subjects recognize. Indeed, most aspects of aggregate behavior are in line with the qualitative predictions of Kamenica and Gentzkow (2011) and our umbrella framework reveals that, indeed, average behavior moves in the direction identified by our comparative statics. However, we also find important differences across subjects that is systematic and can be classified into recognizable paterns of behavior. These reveal that systematic deviations that have been identified in the prior experimental literature on cheap-talk games are a particular type of deviations (over-communication and following messages that should not carry information), but that the opposite types of deviations (under-communication and ignoring meaningful messages) also exist when the setting allows for it.

Overall, the key forces at play in the Kamenica and Gentzkow (2011) model seems to be ones that subjects react to despite subjects not being perfectly rational and optimizing. In this sense, our paper offers a useful starting point to model communication in the presence of commitment, our extension to partial commitment offers a strong experimental device for testing purposes, but also opens an interesting avenue to explore in its own right as partial commitment seems the rule rather than the exception.

---

[54]See also Nguyen (2017) and Au and Li (2018).

References

ABU-MOSTAFA, Y. S., M. MAGDON-ISMAIL, AND H.-T. LIN (2012): *Learning from data*, vol. 4, AMLBook New York, NY, USA:.

ALONSO, R. AND O. CAMARA (2016): "Persuading Voters," *The American Economic Review*, 106, 3590–3605(16).

AU, P. H. AND K. K. LI (2018): "Bayesian Persuasion and Reciprocity Concern: Theory and Experiment," *Working Paper*.

AUSTEN-SMITH, D. (1993): "Information and Influence: Lobbying for Agendas and Votes," *American Journal of Political Science*, 37(3), 799–833.

BAJARI, P. AND A. HORTACSU (2005): "Are Structural Estimates of Auction Models Reasonable? Evidence from Experimental Data," *Journal of Political Economy*, Vol. 113, No. 4, pp. 703–741.

BARDHI, A. AND Y. GUO (2018): "Modes of persuasion toward unanimous consent," *Theoretical Economics*, 13(3), 1111–1149.

BATTAGLINI, M. (2002): "Multiple Referrals and Multidimensional Cheap Talk," *Econometrica*, 70(4), 1379–1401.

BATTIGALLI, P. AND M. SINISCALCHI (2002): "Strong Belief and Forward-Induction Reasoning," *Journal of Economic Theory*.

BENNDORF, V., D. KÜBLER, AND H.-T. NORMANN (2015): "Privacy concerns, voluntary disclosure of information, and unraveling: An experiment," *European Economic Review*, 75, 43–59.

BERGEMANN, D., B. BROOKS, AND S. MORRIS (2015): "The Limits of Price Discrimination," *American Economic Review*, Vol. 105, No. 3, pp. 921–57.

BERGEMANN, D. AND S. MORRIS (2019): "Information Design: A Unified Perspective," *Journal of Economic Literature*.

BLUME, A., D. DE JONG, Y. KIM, AND G. SPRINKLE (1998): "Experimental Evidence on the Evolution of Meaning of Messages in Sender-Receiver Games," *The American Economic Review*, 88, 1323–1340.

BLUME, A., E. K. LAI, AND W. LIM (2017): "Strategic Information Transmission: A Survey of Experiments and Theoretical Foundations," *Working Paper*.

——— (2019): "Mediated Talk: An Experiment," *Working Paper*.

BOCK, O., I. BAETGE, AND A. NICKLISCH (2014): "hroot: Hamburg registration and organization online tool," *European Economic Review*, 71, 117–120.

CAI, H. AND J. T. Y. WANG (2006): "Overcommunication in strategic information transmission games," *Games and Economic Behavior*, 56, 7–36.

CAMERER, C. (1998): "Bounded rationality in individual decision making," *Experimental economics*, 1, 163–183.

CAMERER, C., S. NUNNARI, AND T. R. PALFREY (2016): "Quantal Response and Nonequilibrium Beliefs Explain Overbidding in Maximum-Value Auctions," *Games and Economic Behavior*, Vol 98, 243–263.

CAMERON, A. C. AND D. L. MILLER (2015): "A practitioner's guide to cluster-robust inference," *Journal of Human Resources*, 50, 317–372.

CARTER, A. V., K. T. SCHNEPEL, AND D. G. STEIGERWALD (2017): "Asymptotic behavior of at test robust to cluster heterogeneity," *Review of Economics and Statistics*.

CHARNESS, G. AND D. LEVIN (2005): "When optimal choices feel wrong: A laboratory study of Bayesian updating, complexity, and affect," *American Economic Review*, 95, 1300–1309.

CRAWFORD, V. P. AND J. SOBEL (1982): "Strategic Information Transmission," *Econometrica*, 50, 1431–1451.

DICKHAUT, J., M. LEDYARD, A. MUKHERJI, AND H. SAPRA (2003): "Information management

and valuation: an experimental investigation," *Games and Economic Behavior*, 44, 26–53.

DICKHAUT, J., K. MCCABE, AND A. MUKHERJI (1995): "An Experimental Study of Strategic Information Transmission," *Economic Theory*, 6, 389–403.

DRANOVE, D. AND G. JIN (2010): "Quality Disclosure and Certification: Theory and Practice," *Journal of Economic Literature*, 48, 935–963.

DUFFIE, D., P. DWORCZAK, AND H. ZHU (2017): ""Benchmarks in Search Markets." *Journal of Finance*, 72 (5): 1983–2044.

DYE, R. (1985): "Disclosure of Nonproprietary Information," *Journal of Accounting Research*, 23(1), 123–145.

EMBREY, M., G. R. FRÉCHETTE, AND S. YUKSEL (2017): "Cooperation in the finitely repeated prisoner's dilemma," *The Quarterly Journal of Economics*, 133, 509–551.

EPSTEIN, L. G. AND Y. HALEVY (2019): "Hard-to-Interpret Signals," *Working Paper*.

FORSYTHE, R., R. M. ISAAC, AND T. R. PALFREY (1989): "Theories and Tests of "Blind Bidding" in Sealed-bid Auctions," *The RAND Journal of Economics*, 20, 214–238.

FORSYTHE, R., R. LUNDHOLM, AND T. RIETZ (1999): "Cheap Talk, Fraud, and Adverse Selection in Financial Markets: Some Experimental Evidence," *The Review of Financial Studies*, 12, 481–518.

FRÉCHETTE, G. R. (2012): "Session-Effects in the Laboratory," *Experimental Economics*, 15, 485–498.

GALOR, E. (1985): "Information Sharing in Oligopoly," *Econometrica*, 53, 329–343.

GALPERTI, S. (2019): "Persuasion: The Art of Changing Worldviews," *American Economic Review*,, Vol. 109, No. 3, pp. 996–1031.

GENTZKOW, M. AND E. KAMENICA (2014): "Costly Persuasion," *American Economic Review*, 104, 457–462.

GILLIGAN, T. W. AND K. KREHBIEL (1987): "Decisionmaking and Standing Committees: An Informational Rationale for Restrictive Amendment Procedures," *Journal of Law, Economics*, 3, 287–335.

——— (1989): "Information and Legislative Rules with a Heterogeneous Committee," *American Journal of Political Science*, 33, 459–490.

GOEREE, J. K., C. A. HOLT, AND T. R. PALFREY (2016): *Quantal Response Equilibrium: A Stochastic Theory of Games*, Princeton University Press.

GROSSMAN, S. J. (1981): "The Informational Role of Warranties and Private Disclosure about Product Quality," *The Journal of Law and Economics*, 24, 461.

HAGENBACH, J., F. KOESSLER, AND E. PEREZ-RICHET (2014): "Certifiable Pre-Play Communication: Full Disclosure," *Econometrica*, 82(3), 1093–1131.

HAGENBACK, J. AND E. PEREZ-RICHET (2018): "Communication with Evidence in the Lab," *Working Paper*.

HART, S., I. KREMER, AND M. PERRY (2017): "Evidence Games: Truth and Commitment," *American Economic Review*.

HASTIE, T., R. TIBSHIRANI, AND J. FRIEDMAN (2009): *The elements of statistical learning: data mining, inference, and prediction*, Springer series in statistics New York, NY, USA:, second Edition.

HOLT, C. A. (2007): *Markets, games, and strategic behavior*, Pearson Addison Wesley Boston, MA.

IBRAGIMOV, R. AND U. K. MÜLLER (2010): "t-Statistic based correlation and heterogeneity robust inference," *Journal of Business & Economic Statistics*, 28, 453–468.

JIN, G. AND P. LESLIE (2003): "The effect of information on product quality: Evidence from restaurant hygiene grade cards," *The Quarterly Journal of Economics*.

JIN, G., M. LUCA, AND D. MARTIN (2016): "Is no news (perceived as) bad news? An experimental

investigation of information disclosure," *NBER Working Paper*.

JOVANOVIC, B. (1982): "Truthful Disclosure of Information," *Bell Journal of Economics*, 13, 36–44.

KAMENICA, E. AND M. GENTZKOW (2011): "Bayesian persuasion," *American Economic Review*, 101, 2590–2615.

KAWAGOE, T. AND H. TAKIZAWA (2009): "Equilibrium Refinement vs Level-k Analysis: An Experimental Study of Cheap-Talk Games with Private Information," *Games and Economic Behavior*.

KHALMETSKIA, K., B. ROCKENBACH, AND P. WERNER (2017): "Evasive lying in strategic communication," *Journal of Public Economics*, Vol 156, pp. 59–72.

KING, R. AND D. WALLIN (1991): "Market-induced information disclosures: An experimental markets investigation ," *Contemporary Accounting Research*, 8, 170–197.

LIPNOWSKI, E., D. RAVID, AND D. SHISHKIN (2018): "Persuasion via Weak Institutions," *Working Paper*.

MACQUEEN, J. (1967): "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Oakland, CA, USA, vol. 1, 281–297.

MATHIOS, A. (2000): "The impact of mandatory disclosure laws on product choices: An analysis of the salad dressing market," *The Journal of Law and Economics*.

MILGROM, P. (1981): "Good News and Bad News: Representation Theorems and Applications," *The Bell Journal of Economics*, 12, 380–391.

MIN, D. (2017): "Bayesian Persuasion under Partial Commitment," *Working Paper*.

MURPHY, K. P. (2012): *Machine learning : a probabilistic perspective*, MIT Press.

NAGEL, R. (1995): "Unraveling in Guessing Games: An Experimental Study," *American Economic Review*.

NGUYEN, Q. (2017): "Bayesian Persuasion: Evidence from the Laboratory," *Working Paper*.

OKUNO-FUJIWARA, M., A. POSTLEWAITE, AND K. SUZUMURA (1990): "Strategic Information Revelation," *The Review of Economic Studies*, 57, 25–47.

SÁNCHEZ-PAGÉS, S. AND M. VORSATZ (2007): "An experimental study of truth-telling in a sender-receiver game," *Games and Economic Behavior*, 61, 86–112.

VERRECCHIA, R. E. (1983): "Discretionary Disclosure." *Journal of Accounting and Economics*, 5, 179–194.

WANG, J., M. SPEZIO, AND C. CAMERER (2010): "Pinocchio's Pupil: Using Eyetracking and Pupil Dilation to Understand Truth Telling and Deception in Sender-Receiver Games," *The American Economic Review*, 100, 984–1007.

WILSON, A. AND E. VESPA (2017): "Information Transmission Under the Shadow of the Future: An Experiment," *Working Paper*.

Guillaume Fréchette Alessandro Lizzeri Jacopo Perego
NYU   NYU   Columbia

Contents

A   Equilibrium, Refinements and Proofs

A.1   Equilibrium Characterization

In this section, we characterize the set of Perfect Bayesian Equilibria (PBE) for the framework introduced in Section 3. In a PBE, the sender chooses an information structure $\pi_C \in \Pi$ in the commitment stage. every history π'_C, the sender chooses $\pi'_R \in \Pi$, possibly as a function of π'_C. For notational simplicity, we omit the dependence of π'_R on π'_C. The receiver observes history (m, π_C) and responds with an action in $\{a_H, a_L\}$. We call such an action $a(m, \pi_C)$. Finally, a belief assessment μ assigns a belief to every triple (m, π'_C, π'_R).

Definition A1. *Fix (Π, ρ, q). The tuple (π_C, π_R, a, μ) is a Perfect Bayesian Equilibrium if:*

(1) *π_C maximizes $\sum_{\theta, m} \mu_0(\theta)\big(\rho\pi_C(m|\theta) + (1 - \rho)\pi_R(m|\theta)\big)v(a(m, \pi_C))$;*

(2) *For all (π'_C, θ), π'_R maximizes $\sum_m \pi'_R(m|\theta)v\big(a(m, \pi'_C)\big)$;*

(3) *For all (m, π'_C), $a(m, \pi'_C) = a_H$ if $\mu(m, \pi'_C, \pi'_R) \geq q$;*

(4) *For all (m, π'_C, π'_R), posterior belief $\mu(m, \pi'_C, \pi'_R)$ is computed from $\pi := \rho\pi'_C + (1 - \rho)\pi'_R$ using Bayes' rule whenever possible.*

Next, we provide a characterization of the equilibrium set, before imposing further qualification on our notion of equilibrium. For this result, we say that an equilibrium (π_C, π_R, a, μ) under (Π, ρ, q) achieves *full-commitment informativeness* if $\phi^B(\rho\pi_C + (1 - \rho)\pi_R) = (\frac{q-\mu_0}{1-\mu_0})^{\frac{1}{2}}$. We label such equilibria as FCI. This is the equilibrium informativeness under full commitment and unverifiable information, which is an important benchmark. It is also useful to define two thresholds for ρ: $\underline{\rho} := \frac{q-\mu_0}{q(1-\mu_0)}$ and $\bar{\rho} = \frac{q(1-\mu_0)}{q(1-\mu_0)+(1-q)\mu_0}$ and note that $\underline{\rho} \leq \bar{\rho}$. We first consider the case of unverifiable information.

Proposition A1. *Fix $q > \mu_0$ and assume that information be unverifiable.*

(a) *If $\rho < \underline{\rho}$, then all equilibria are uninformative.*

(b) *If $\rho \in [\underline{\rho}, \bar{\rho})$, then there exist FCI equilibria and uninformative equilibria. There also exist equilibria that are more informative than FCI.*

(c) *If $\rho \geq \bar{\rho}$, there exist FCI equilibria. There is no uninformative equilibrium. There also exist equilibria that are more informative than FCI.*

When commitment power is low, the sender cannot successfully transmit information to the receiver. When commitment power is sufficiently high, all equilibria

involve some information transmission. Furthermore, when commitment power is sufficiently high, the sender can achieve BP-informativeness in equilibrium despite the lack of full commitment. She achieves this outcome by appropriately over-communicating in the commitment stage, correctly anticipating that her own behavior in the revision stage will reduce the credibility of her communication. Next, we turn to the equilibrium characterization when information is verifiable.

Proposition A2. *Fix $q > \mu_0$ and assume that information is verifiable.*

(a) *If $\rho < \underline{\rho}$, all equilibria are fully informative.*

(b) *If $\rho \in [\underline{\rho}, \bar{\rho})$, the least informative equilibrium is FCI; fully informative equilibria exist.*

(c) *If $\rho \geq \bar{\rho}$, there are no fully informative equilibria; the least informative equilibrium is FCI.*

From this proposition, we can appreciate the contrast that information verifiability imposes on the equilibrium set. First, when commitment power is sufficiently low, all equilibria are fully informative, in stark contrast with the unverifiable case. Second, when the commitment is sufficiently high, the sender can avoid the unattractive scenario where she fully disclose her private information. Namely, there are no fully informative equilibria. Third, BP-informativeness can be achieved in equilibrium as long as $\rho \geq \underline{\rho}$. This last property is in line with the unverifiable case.

A.2   Truth-Leaning Equilibria

The analysis above provides a complete characterization of the equilibrium set. In this section, we provide two examples, one for unverifiable and one for verifiable information, of PBEs that do not satisfy the truth-leaning tie-breaking rule that we introduced in Section 3. We use these examples to argue that equilibria that are not truth-leaning feature behavior in the revision stage that is somewhat unreasonable.

Example 1: *Unverifiable Information.*
Assume that information is unverifiable and set the degree of commitment to $\rho = \frac{3}{5}$, the persuasion threshold to $q = \frac{1}{2}$, and the prior to $\mu_0 = \frac{1}{3}$. Consider the pair (π_C, π_R) that is reported in Table A7. First note that $\mu(\theta_H, \pi_C, \pi_R) < q$ and $\mu(\theta_L, \pi_C, \pi_R) < q$. That is, despite the fact that π_C is fully informative, the sender's behavior in the revision stage entirely garbles the information from the commitment stage.

How do we support this equilibrium? Suppose that for any deviation at the commitment stage π_C', the sender chooses an appropriate π_R' at the revision stage so as to make the pair (π_C', π_R') uninformative. The Proof of Proposition A1.(b)

2

Table A7

π_C	θ_H	θ_L	n	π_R	θ_H	θ_L	n
θ_H	1	0	0	θ_H	0	1	0
θ_L	0	1	0	θ_L	1	0	0

establishes that, for ρ sufficiently low, such a π'_R exists. Given the receiver's beliefs about the revision stage strategy, the receiver would choose action a_L for both messages. Thus, the sender is indifferent among all her strategies in the revision stage and is willing to choose π'_R. Furthermore, given the receiver's expectation about π'_R, in the commitment stage the sender is also indifferent among all his strategies: all of them lead to a payoff of zero. This particularly strange behavior of the sender in the revision stage is ruled out by our tie-breaking rule. In this equilibrium, in the revision stage the sender of type θ_H is indifferent between sending message θ_L and being truthful. Truth-leaning requires that such a sender choose $\pi_R(\theta_H|\theta_L) = 1$ instead.

Example 2: *Verifiable Information.*

Now assume that information is verifiable. As above, we set the degree of commitment to $\rho = \frac{3}{5}$, the persuasion threshold to $q = \frac{1}{2}$ and the prior to $\mu_0 = \frac{1}{3}$. We consider the pair (π_C, π_R) that is described in Table A8.

Table A8

π_C	θ_H	θ_L	n	π_R	θ_H	θ_L	n
θ_H	0	0	1	θ_H	0	0	1
θ_L	0	$\frac{5}{6}$	$\frac{1}{6}$	θ_L	0	0	1

Given π_C as in the table, in the revision stage the sender of type θ_L strictly prefers message n to message θ_L, whereas the sender of type θ_H is indifferent among the two feasible messages. Furthermore, it can be verified that the pair (π_C, π_R) described in the table is FCI, i.e. it leads to the maximal achievable equilibrium payoff for the sender at the commitment stage. Therefore, the sender has no incentive to deviate at the commitment stage. This equilibrium relies on implausible behavior in the revision stage. To see this, consider the on-path decision of the sender of type θ_H in the revision stage. She can choose between sending message n, which induces a belief of $\frac{1}{2}$, or sending an off-path message θ_H, which induce a belief of 1. The former leaves the receiver indifferent, the latter doesn't. While consistent with the requirement of PBE, this equilibrium is not truth-leaning because the sender is not telling the whole truth even if it is optimal to do so.

3

Despite being a simple tie-breaking rule that applies only to a single type of sender, this refinement is powerful enough to select a unique equilibrium outcome for each combination of ρ, q and μ_0. The following result formalizes this intuition.

Proposition A3. *Fix $\rho \in [0,1]$ and $q > \mu_0$.*

> *(Unverifiable) If $\rho < \underline{\rho}$, truth-leaning equilibria are uninformative. If $\rho \geq \underline{\rho}$, truth-leaning equilibria are FCI.*

> *(Verifiable) If $\rho < \bar{\rho}$, truth-leaning equilibria are fully informative. If $\rho \geq \bar{\rho}$, all truth-leaning equilibria are equally informative.*

A.3 Proofs

A.3.1 Proof of Proposition A1

Proof of Proposition A1.(a). Let information be unverifiable and $\rho < \underline{\rho}$. Suppose by way of contradiction that there is an equilibrium (π_C, π_R, a, μ) such that $\phi^B(\rho \pi_C + (1-\rho)\pi_R) > 0$. This implies that there are messages realizing with positive probability that lead to action a_H. There are two cases to consider.

Case 1. There exists exactly one message m' realizing with positive probability such that $a(m, \pi_C, \pi_R) = a_H$. When this case, equilibrium conditions imply that $\pi_R(m'|\theta) = 1$ for all θ. However, given this we have that

$$
\begin{aligned}
q \leq \mu(m') &= \frac{\mu_0(\rho\pi_C m'|\theta_H) + (1-\rho))}{\mu_0(\rho\pi_C(m'|\theta_H) + (1-\rho)) + (1-\mu_0)(\rho\pi_C(m'|\theta_L) + (1-\rho))} \\
&\leq \frac{\mu_0}{\mu_0 + (1-\mu_0)(1-\rho)} \\
&< \frac{\mu_0}{\mu_0 + (1-\mu_0)(1-\underline{\rho})} = q.
\end{aligned}
$$

The first inequality holds since m' leads to action a_H. The first equality follows from Bayes' rule. The second inequality, instead, hold because $\mu(m')$ is maximized when we set $\pi_C(m'|\theta_H) = 1 - \pi_C(m'|\theta_L) = 1$. The third inequality, instead, holds because $\rho < \underline{\rho}$. This leads to a contradiction, and therefore we can rule out Case 1.

Case 2. There are exactly two messages $m', m'' \in M$ realizing with positive probability such that $a(m, \pi_C, \pi_R) = a_H$, for $m \in \{m', m''\}$. Denote $\pi_i(m', m''|\theta) := \pi_i(m'|\theta) + \pi_i(m''|\theta)$, for all θ and $i \in \{C, R\}$. Because both m' and m'' lead to a_H, equilibrium conditions imply that $\pi_R(m', m''|\theta) = 1$ for all θ. Denote $\mu(m', m'')$ the

posterior belief conditional on observing m' or m''. That is,

$$
\begin{aligned}
\mu(m', m'') &= \frac{\mu_0(\rho(\pi_C(m', m''|\theta_H) + (1 - \rho))}{\mu_0\rho\pi_C(m', m''|\theta_H) + (1 - \mu_0)\rho\pi_C(m', m''|\theta_L) + (1 - \rho)} \\
&\leq \frac{\mu_0}{\mu_0 + (1 - \mu_0)(1 - \rho)} \\
&< \frac{\mu_0}{\mu_0 + (1 - \mu_0)(1 - \underline{\rho})} = q.
\end{aligned}
$$

The first inequality holds because $\mu(m', m'')$ is maximized when $\pi_C(m', m''|\theta_H) = 1 - \pi_C(m', m''|\theta_L) = 1$. This shows that $\mu(m', m'') < q$. However, Bayes' rule also implies that, for appropriately chosen weight β,[55]

$$
\mu(m', m'') = \beta\mu(m') + (1 - \beta)\mu(m'') \geq q.
$$

Therefore, we have $q \leq \mu(m', m'') < q$, a contradiction. We conclude that the equilibrium cannot be informative. □

Proof of Proposition A1.(b).

Existence of BP-informative equilibria.

Fix $\rho \geq \underline{\rho}$. We first show that for BP-informativeness can be achieved in equilibrium. We do so by constructing such equilibrium. We start by defining strategies on the equilibrium path. For the commitment stage, let $\pi_C(m'|\theta_H) = 1$, $\pi_C(m'|\theta_L) = x$ and $\pi_C(m''|\theta_L) = 1 - x$, where $x = \frac{1}{\rho}\left(\frac{\mu_0(1-q)}{q(1-\mu_0)} - (1 - \rho)\right)$. Note that π_C is well-defined. On the one hand, $x \geq 0$ if $\frac{\mu_0(1-q)}{q(1-\mu_0)} \geq 1 - \underline{\rho} \geq 1 - \rho$, which is true since $1 - \underline{\rho} = \frac{\mu_0(1-q)}{q(1-\mu_0)}$. On the other hand, $x \leq 1$ follows directly from our maintained assumption $q > \mu_0$. For the revision stage, let $\pi_R(m'|\theta) = 1$, for all θ. Given this choice of π_C and π_R, we have that $\mu(m', \pi_C, \pi_R) = q$ and $\mu(m'', \pi_C, \pi_R) = 0$, hence let $a(m', \pi_C) = a_H$ and $a(m'', \pi_C) = a_H$. It is straightforward to check that $\phi^B(\rho\pi_C + (1-\rho)\pi_R) = \left(\frac{q-\mu_0}{1-\mu_0}\right)^{\frac{1}{2}}$, the BP-informativeness. We now define strategies off the equilibrium path. For any π'_C, let $\mu(m, \pi'_C) = \frac{\mu_0\pi'_C(m|\theta_H)}{\mu_0\pi'_C(m|\theta_H) + (1-\mu_0)\pi'_C(m|\theta_L)}$. Let $\bar{m}$ be such that $\mu(\bar{m}, \pi'_C) \geq \mu(m, \pi'_C)$, for all $m \in M$. Let $\pi'_R(\bar{m}|\theta) = 1$ for all θ. For thus defined pairs (π'_C, π'_R), let $a(m, \pi'_C) = a_H$ if and only if $\mu(m, \pi'_C, \pi'_R) \geq q$. Whenever a message m has zero probability let $\mu(m, \pi'_C, \pi'_R) = 0$. It is straightforward to check that this strategy is indeed an equilibrium and, as noted above, achieves BP-informativeness.

Existence of uninformative equilibria.

Next, we show that when $\rho \in [\underline{\rho}, \bar{\rho})$, an uninformative equilibrium exists. The proof is by construction and consists in finding, for each possible history π_C, a revision strategy π_R such that $\phi^B(\rho\pi_C + (1 - \rho)\pi_R) = 0$. The existence of such π_R for

---

[55]More specifically, $\beta := \frac{\sum_\theta \mu_0(\theta)(\rho\pi_C(m'|\theta) + (1-\rho)\pi_R(m'|\theta))}{\sum_\theta \mu_0(\theta)(\rho\pi_C(m', m''|\theta) + (1-\rho)\pi_R(m', m''|\theta))}$

each history π_C guarantees the existence of an uninformative equilibrium. To this purpose, consider an arbitrary π_C. If $\mu(m, \pi_C) < q$, for all $m \in M$, then let $\pi_R = \pi_C$, which gives $\phi^B(\rho\pi_C + (1-\rho)\pi_R) = 0$. Conversely, suppose that there exists a message $\mu(m, \pi_C) \geq q$. For arbitrary π_C, Bayes plausibility requires that there exists at least one message, call it m''', such that $\mu(m''', \pi_C) \leq \mu_0$. To simplify notation, let $\pi_C(m'|\theta_H) = a'$, $\pi_C(m''|\theta_H) = a''$, $\pi_C(m'''|\theta_H) = a'''$, $\pi_C(m'|\theta_L) = b'$, $\pi_C(m''|\theta_L) = b''$, $\pi_C(m'''|\theta_L) = b'''$. Define the revision strategy as follows: $\pi_R(m'''|\theta_H) = 1$, and let $\pi_R(m'|\theta_L) = x'$, $\pi_R(m''|\theta_L) = x''$ and $\pi_R(m'''|\theta_L) = x'''$. We want to show that there exists (x', x'', x''') such that $x' + x'' + x''' = 1$ and $\pi(m, \pi_C, \pi_R) < q$, for all $m \in M$. We have that $\pi(m', \pi_C, \pi_R) < q$ is equivalent to:

$$x' > \Phi' := \frac{\rho}{1-\rho}\Big((1-\underline{\rho})a' - b'\Big).$$

Similarly, $\pi(m'', \pi_C, \pi_R) < q$ is equivalent to:

$$x'' > \Phi'' := \frac{\rho}{1-\rho}\Big((1-\underline{\rho})a'' - b''\Big).$$

Finally, the last condition $\pi(m''', \pi_C, \pi_R) < q$ is equivalent to:

$$x' + x'' < \bar{\Phi} := \underline{\rho} + \frac{\rho}{1-\rho}\Big(b''' - a''' + \underline{\rho}a'''\Big).$$

It is straightforward to check that $\Phi' + \Phi'' < \bar{\Phi}$ and also that $\Phi' + \Phi'' < 1$ if and only if $\rho < \bar{\rho}$. Therefore, x' and x'' can be found so that the thus defined π_R is an information structure and $\phi^B(\rho\pi_C + (1-\rho)\pi_R) = 0$.

It is straightforward to complete the construction of the uninformative equilibrium. Note that the sender has not profitable deviation in the commitment stage. In fact, all possible deviations π'_C lead to a π'_R that, by construction, only induces beliefs strictly below q, hence a guess a_L. Similarly, the sender has no profitable deviation in the revision stage, for reasons that are similar to the existence of a babbling equilibrium in a cheap talk game.

Existence of equilibria that are more than BP-informative.

The construction of this equilibrium is tightly related to the construction of uninformative equilibria above. Fix $\rho \geq \underline{\rho}$. We start by constructing the sender's strategies on the equilibrium path. Let $\pi_C(m'|\theta_H) = \pi_C(m''|\theta_L) = 1$, that is π_C is fully informative. Let $\pi_R(m'|\theta) = 1$ for all θ. Following these choices, beliefs and receiver's guesses are naturally pinned down. For all "off-path" $\pi'_C \neq \pi_C$, we associate a π'_R that is constructed as in the case of an uninformative equilibrium, as explained above. This means that for all $\pi'_C \neq \pi_C$, $\pi^B(\rho\pi'_C + (1-\rho)\pi'_R) = 0$, the receiver always guess a_L and the sender's expected utility is 0. Clearly, in light of this construction, the sender in the commitment stage has no incentive to deviate

away from π_C. This thus define an equilibrium. Moreover, it's easy to verify that $\phi^B(\rho\pi_C + (1-\rho)\pi_R) = \left(\frac{\mu_0\rho}{1-\rho(1-\mu_0)}\right)^{\frac{1}{2}}$, which is higher than BP-informativeness for all $\rho \geq \underline{\rho}$. $\square$

Proof of Proposition A1.(c).
The existence of BP-informative equilibria as well as the existence of equilibria that are more than BP-informative follows directly from the Proof of Proposition A1.(b). For these two results, we didn't use the assumption $\rho < \bar{\rho}$, and therefore they hold for this case as well.

Non-existence of uninformative equilibria.
We now prove that when $\rho \geq \bar{\rho}$ all equilibria are strictly informative. Suppose not. That is let $\rho \geq \bar{\rho}$ and let (π_C, π_R, μ, a) be an uninformative equilibrium. As such, the sender earns a payoff of zero. We construct a profitable deviation π'_C under which, there exists a message m' that induces action a_H with strictly positive probability. We construct this deviation to be fully informative, namely, $\pi'_C(m'|\theta_H) = 1$ and $\pi'_C(m''|\theta_L) = 1$, for $m'' \neq m'$. Call π'_R the continuation strategy of the sender in the revision stage. We have that,

$$\mu(m', \pi'_C, \pi'_R) = \frac{\mu_0(\rho + \pi'_R(m'|\theta_H))}{\mu_0(\rho + \pi'_R(m'|\theta_H)) + (1-\mu_0)(1-\rho)\pi'_R(m'|\theta_L)}$$

$$\geq \frac{\mu_0\rho}{\mu_0\rho + (1-\mu_0)(1-\rho)} \geq \frac{\mu_0\bar{\rho}}{\mu_0\bar{\rho} + (1-\mu_0)(1-\bar{\rho})} = q$$

The first inequality holds because setting $\pi'_R(m'|\theta_H) = 0$ and $\pi'_R(m'|\theta_L) = 1$ induces a lower bound for $\mu(m', \pi'_C, \pi'_R)$. The second inequality holds because $\rho \geq \bar{\rho}$, by assumption. Therefore, in the continuation game following deviation π'_C, the receiver plays a_H with positive probability and the sender earns a strictly positive payoff. Hence, the deviation is strictly profitable. $\square$

A.3.2 Proof of Proposition A2

Proof of Proposition A2.(a). Let information be verifiable and $\rho < \underline{\rho}$. We want to show that all equilibria are fully informative. Suppose not, that is (π_C, π_R, a, μ) is an equilibrium with $\phi^B(\rho\pi_C + (1-\rho)\pi_R) < 1$. Because information is verifiable, such a situation must imply that $a(n, \pi_C) = a_H$. Therefore, $\mu(n, \pi_C, \pi_R) \geq q$. However, equilibrium conditions also imply that $\pi_R(n|\theta_L) = 1$, i.e. in the revision stage, the sender of type θ_L always sends message n. Therefore, we have that:

$$q \leq \mu(n, \pi_C, \pi_R) \leq \frac{\mu_0}{\mu_0 + (1-\mu_0)(1-\rho)} < \frac{\mu_0\underline{\rho}}{\mu_0\underline{\rho} + (1-\mu_0)(1-\underline{\rho})} = q.$$

The first inequality comes from $a(n, \pi_C) = a_H$. The second inequality holds because setting $\pi_C(n|\theta_H) = 1$, $\pi_C(n|\theta_L) = 0$ and $\pi_R(n|\theta_R) = 1$ generates an upper bound for the value of $\mu(n, \pi_C, \pi_R)$. The last inequality holds because $\rho < \underline{\rho}$, by assumption. Therefore, $q \leq \mu(n, \pi_C, \pi_R) < q$, a contradiction. $\qquad\square$

Proof of Proposition A2.(b).

Existence of BP-informative equilibria.

We prove this by construction. Fix $\rho \geq \underline{\rho}$. Let π_C and π_R be such that $\pi_C(n|\theta_H) = \pi_R(n|\theta_H) = \pi_R(n|\theta_L) = 1$ and $\pi_C(n|\theta_L) = x$. Let $x := \rho - \underline{\rho} \geq 0$, by assumption. Moreover, it is easy to verify that $\mu(n, \pi_C, \pi_R) = q$. Let $a(n, \pi_C) = a_H$, therefore π_R is a best response to π_C given the receiver behavior. It is also easy to verify $\phi^B(\rho\pi_C + (1-\rho)\pi_R) = (\frac{q-\mu_0}{1-\mu_0})^{\frac{1}{2}}$. That is (π_C, π_R) is BP-informative. As a consequence, no profitable deviation away from π_C exists. Therefore, we have have constructed an equilibrium that is BP-informative. Moreover, this is the least informative equilibrium that exists in this case. To see this, note that, because of the nature of verifiable information, $\phi^B(\rho\pi_C + (1-\rho)\pi_R) < 1$ requires that $\mu(n, \pi_C, \pi_R) \in [q, 1)$. Moreover, $\phi^B(\rho\pi_C + (1-\rho)\pi_R)$ is increasing in $\mu(n, \pi_C, \pi_R)$. The equilibrium that we constructed above has $\mu(n, \pi_C, \pi_R) = q$ and it is therefore minimally informative.

Existence of fully informative equilibria.

We prove this by construction. Consider a revision strategy π_R defined as $\pi_R(\theta_H|\theta_H) = \pi_R(n|\theta_L) = 1$. Moreover, suppose that π_R is played for all histories π_C'. Consider an arbitrary history π_C'. Note that, for all $\rho < \bar{\rho}$,

$$\mu(n, \pi_C', \pi_R) = \frac{\mu_0 \rho \pi_C'(n|\theta_H)}{\mu_0 \rho \pi_C'(n|\theta_H) + (1-\mu_0)(\rho \pi_C'(n|\theta_L) + (1-\rho))} < q.$$

Moreover, note that π_R is a best-response to this arbitrary π_C'. Finally, note that in the subgame indexed by π_C', the sender expects to receive a payoff of $\mu_0(\rho\pi_C'(\theta_H|\theta_H) + (1-\rho)\pi_R(\theta_H|\theta_H) \leq \mu_0$. Now consider the strategy $\pi_C = \pi_R$. This strategy gives a payoff of μ_0 and, thanks to the argument above, no profitable deviation away from this strategy exists. Moreover, $\phi^B(\rho\pi_C' + (1-\rho)\pi_R') = 1$. $\quad\square$

Proof of Proposition A2.(c). The existence of BP-informative equilibria as well as the fact that these are the least informative equilibria follows directly from the Proof of Proposition A2.(b). For proving these two results, we didn't use the assumption $\rho < \bar{\rho}$, and therefore the same proof hold for the case $\rho \geq \bar{\rho}$ as well.

Non-existence of fully informative equilibria.

We first show that when $\rho \geq \bar{\rho}$, there exists no fully informative equilibrium. When $\rho = 1$ the result is a straightforward consequence of full commitment, so let's focus on the case $\rho \in [\bar{\rho}, 1)$. Suppose that there exists an equilibrium (π_C, π_R, a, μ) such

that $\phi^B(\rho\pi_C + (1-\rho)\pi_R) = 1$. In this equilibrium, the sender expects to earn μ_0. Consider a deviation π'_C such that $\pi'_C(n|\theta_H) = 1$ and $\pi'_C(\theta_L|\theta_L) = 1$. We argue that this deviation leads to a subgame in which the sender earns strictly more than μ_0. First, note that for all π'_R,

$$\mu(n, \pi'_C, \pi'_R) = \frac{\mu_0(\rho + (1-\rho)\pi'_R(n|\theta_H))}{\mu_0(\rho + (1-\rho)\pi'_R(n|\theta_H)) + (1-\mu_0)(1-\rho)\pi'_R(n|\theta_L)} \geq$$

$$\geq \frac{\mu_0\rho}{\mu_0\rho + (1-\mu_0)(1-\rho)} \geq \frac{\mu_0\bar{\rho}}{\mu_0\bar{\rho} + (1-\mu_0)(1-\bar{\rho})} = q.$$

Therefore, $a(\pi'_C, n) = a_H$. This implies that $\pi'_R(n|\theta_L) = 1$. Hence, the expected payoff for the sender in the commitment stage is bounded below by $\mu_0\pi'_C(n|\theta_H) + (1-\mu_0)\pi'_R(n|\theta_L) = 1 > \mu_0$. Therefore, π'_C is a profitable deviation. Moreover, irrespective of what $\pi'_R(n|\theta_H)$ is, the fact that n is sent with strictly positive probability from both states implies, as long as $\rho < 1$, that $\mu(n, \pi'_C, \pi'_R) < 1$, hence $\phi^B(\rho\pi'_C + (1-\rho)\pi'_R) < 1$. $\qquad\square$

A.3.3 Proof of Proposition A3.

Unverifiable Information. If $\rho < \underline{\rho}$, we know by Proposition A1.(a) that all PBE are uninformative. A fortiori, all truth-leaning are uninformative. Note that, truth-leaning equilibria exists in this case. For example, let π_C and π_R be defined as $\pi_C(\theta|\theta) = \pi_R(\theta_H|\theta) = 1$ for all θ, $\mu(m, \pi_C, \pi_R) = \mu_0$, and $a(m, \pi_C) = a_L$. Therefore, consider instead the case $\rho \geq \underline{\rho}$. We want to argue that all TLE are BP-informative. In order to do so, we argue that there exists a pair (π_C, π_R) such that (1) π_R is a best-response to π_C, (2) π_R is uniquely pinned down by the truth-leaning refinement and, moreover, (3) $\phi^B(\rho\pi_C + (1-\rho)\pi_R) = (\frac{q-\mu_0}{1-\mu_0})^{\frac{1}{2}}$. To this purpose, let $\pi_C(\theta_H|\theta_H) = 1$, $\pi_C(\theta_H|\theta_L) = x$ and $\pi_C(\theta_L|\theta_L) = 1 - x$, where $x := \frac{1}{\rho}(\rho - \underline{\rho})$. Note that $x \in [0, 1]$, hence π_C is well-defined. Conversely, let π_R be such that $\pi_R(\theta_H|\theta) = 1$ for all θ. First, let's establish that π_R is best-responding to π_C. To see this note that, by construction, $\mu(\theta_L, \pi_C, \pi_R) = 0$ and $\mu(\theta_H, \pi_C, \pi_R) = q$. Therefore, $a(\theta_H, \pi_C) = a_H$ and $a(\theta_L, \pi_C) = a_L$. Consistently, π_R gives positive probability to $m = \theta_H$ only. Hence π_R is best-responding to π_C. Second, let's argue that π_R is indeed truth-leaning. To see this, just notice that, in the revision stage, the sender of type θ_H is being truthful, hence π_R is truth-leaning. (In passing, note that also type θ_L is truth-leaning since she is not indifferent between $m = \theta_H$ and $m = \theta_L$). Finally, it is straightforward to verify that, given this choice of (π_C, π_R), we have that $\phi^B(\rho\pi_C + (1-\rho)\pi_R) = (\frac{q-\mu_0}{1-\mu_0})^{\frac{1}{2}}$. Note that this is the BP-informativeness, namely the equilibrium informativeness when the sender has full commitment power. This implies that, if the pair (π_C, π_R) is played on the equilibrium path, it leads to

9

the first-best payoff, namely $\frac{\mu}{q}$. This proves that all truth-leaning equilibria of the grand-game are BP-informative. To see this, suppose that this is not the case, i.e. there exists a truth-leaning equilibrium $(\pi'_C, \pi'_R, \mu', a')$ that is not BP-informative. Because of this, sender's expected payoff in this equilibrium is strictly smaller than $\frac{\mu}{q}$. However, a deviation at the commitment stage exists, namely strategy π_C, that leads to a unique best-response in the revision stage, namely π_R, that is consistent with truth-leaning and that achieves the first-best payoff, namely $\frac{\mu}{q}$. Therefore, such deviation is strictly profitable and $(\pi'_C, \pi'_R, \mu', a')$ is not an equilibrium.

Verifiable Information. If $\rho < \underline{\rho}$, we know by Proposition A2.(a) that all PBE are fully informative. A fortiori, all truth-leaning equilibria are fully informative. Trivially, a truth-leaning equilibrium exists. For example, $\pi_i(\theta|\theta) = 1$ for all θ and $i \in \{C, R\}$; $\mu(m, \pi_C, \pi_R) = 1$ if $m = \theta_H$ and 0 otherwise; $a(m, \pi_C) = a_H$ iff $m = \theta_H$ and a_L otherwise. Therefore, consider instead the case $\rho \in [\underline{\rho}, \bar{\rho})$. We want to show that all truth-leaning equilibria are fully informative. Suppose not, namely let (π_C, π_R, μ, a) be a truth-leaning equilibrium such that $\phi^B(\rho\pi_C + (1-\rho)\pi_R) < 1$. Since equilibrium informativeness is strictly less than one, there must exist a message m such that $\mu(m, \pi_C, \pi_R) \in (0, 1)$. When information is verifiable, it is necessarily the case that $m = n$. Moreover, $a(n, \pi_C) = a_H$. If this was not the case, $\pi_i(n|\theta_H) = 0$, for $i \in \{C, R\}$, hence $\mu(n, \pi_C, \pi_R) = 0$, a contradiction. Therefore, let $\mu(n, \pi_C, \pi_R) \in [q, 1)$. On the one hand, equilibrium requires that $\pi_R(n|\theta_L) = 1$. (Incidentally, note that this is consistent with truth-leaning since the two messages lead to different payoffs). On the other hand, type θ_H in the revision stage is indifferent between θ_H and n, as they both lead to action a_H. The truth-leaning refinement requires that $\pi_R(\theta_H|\theta_H) = 1$. Therefore, the fact that the equilibrium is not fully informative uniquely pins down π_R. Given this, we note that:

$$\mu(n, \pi_C, \pi_R) \leq \frac{\mu\rho}{\mu\rho + (1-\mu)(1-\rho)} < \frac{\mu\bar{\rho}}{\mu\bar{\rho} + (1-\mu)(1-\bar{\rho})} = q.$$

Hence, $\mu(m, \pi_C, \pi_R) < q$, a contradiction. Finally, let us consider the case $\rho \geq \bar{\rho}$. We want to show that all truth-leaning equilibria are equally informative. Let (π_C, π_R, μ, a) be a truth-leaning equilibrium. By Proposition A2.(c), no equilibrium is fully informative. Therefore, by the same argument made above, $\mu(n, \pi_C, \pi_R) \in [q, 1)$ and π_R is uniquely pinned down. Moreover, π_R is independent of π_C. Therefore, there exists a unique best-response π_C to such revision strategy π_R. Such π_C is given by $\pi_C(n|\theta_H) = 1$ and $\pi_C(n|\theta_L) = x$, where $x := (1 - \underline{\rho}) - \frac{1-\rho}{\rho} \in [0, 1]$. This strategy π_C satisfies $\mu(n, \pi_C, \pi_R) = q$ while maximizing the ex-ante probability of sending message n. By construction, all truth-leaning equilibria share the same on-path sender's behavior (π_C, π_R). Therefore, all truth-leaning equilibria have to be equally informative. Moreover, it is is an exercise in algebra to verify that $\phi^B(\rho\pi_C + (1 - \rho)\pi_R) = \left(\frac{q - \mu_0(\rho + q(1-\rho))}{(1 - \mu_0)(\rho + q(1-\rho))}\right)^{\frac{1}{2}}$. $\qquad \square$

A.3.4 Proof of Proposition 1

In Proposition A3, we have established that, for any given ρ and $q > \mu_0$ and verifiability scenario, all truth-leaning equilibria are equally informative. Now let information be unverifiable. Proposition A3 also establishes that truth-leaning equilibria are uninformative if $\rho < \underline{\rho}$ and BP-informative otherwise. Moreover, $\phi^B(\rho \pi_C + (1-\rho)\pi_R) = \left(\frac{q-\mu_0}{(1-\mu_0)}\right)^{\frac{1}{2}} > 0$, since $q > \mu_0$. Finally, we want to show that, when $\rho \geq \underline{\rho}$, any truth-leaning equilibrium (π_C, π_R, μ, a) satisfies $\phi^B(\pi_C) > \phi^B(\pi_R)$. Since the equilibrium is strictly informative, there exists a message m' inducing action a_H. Then, $\pi_R(m'|\theta) = 1$, for all θ. Therefore, $\phi^B(\pi_R) = 0$. However, $\phi^B(\rho \pi_C + (1-\rho)\pi_R) = \left(\frac{q-\mu_0}{1-\mu_0}\right)^{\frac{1}{2}} > 0$, implying that $\phi^B(\pi_C) > 0$. We conclude that π_C is more informative than π_R. Now let information be verifiable. In Proposition A3, we established that truth-leaning equilibria are fully informative if $\rho < \bar{\rho}$. Moreover, we also established that, if $\rho \geq \bar{\rho}$, any truth-leaning equilibrium (π_C, π_R, μ, a) has $\phi^B(\rho \pi_C + (1-\rho)\pi_R) = \left(\frac{q-\mu_0(\rho+q(1-\rho))}{(1-\mu_0)(\rho+q(1-\rho))}\right)^{\frac{1}{2}} < 1$. In this case, we argued that the fact that truth-leaning equilibria are not fully informative pins down the on-path sender's behavior (π_C, π_R). In particular, we showed that $\pi_C(n|\theta_H) = 1$, $\pi_C(n|\theta_L) = (1-\underline{\rho}) - \frac{1-\rho}{\rho} \in [0,1]$ and that $\pi_R(\theta_H|\theta_H) = \pi_R(n|\theta_L) = 1$. Given this, it is straightforward to conclude that $\phi^B(\pi_C) < \phi^B(\pi_R)$. $\qquad \square$

A.3.5 Proof of Proposition 2

When information is unverifiable, Proposition A3 established that, any truth-leaning equilibrium (π_C, π_R, μ, a) satisfies

$$\phi^B(\rho \pi_C + (1-\rho)\pi_R) = \begin{cases} 0 & \text{if } \rho < \underline{\rho} \\ \left(\frac{q-\mu_0}{1-\mu_0}\right)^{\frac{1}{2}} & \text{if } \rho \geq \underline{\rho} \end{cases}$$

Therefore, when information is unverifiable, equilibrium informativeness is weakly increasing in ρ. Conversely, let information be verifiable. In the proof of Proposition A3, we established that any truth-leaning equilibrium (π_C, π_R, μ, a) satisfies

$$\phi^B(\rho \pi_C + (1-\rho)\pi_R) = \begin{cases} 1 & \text{if } \rho < \bar{\rho} \\ \left(\frac{q-\mu_0(\rho+q(1-\rho))}{(1-\mu_0)(\rho+q(1-\rho))}\right)^{\frac{1}{2}} & \text{if } \rho \geq \bar{\rho} \end{cases}$$

It is easy to verify that $\left(\frac{q-\mu_0(\rho+q(1-\rho))}{(1-\mu_0)(\rho+q(1-\rho))}\right)^{\frac{1}{2}}$ is decreasing (strictly) in ρ. Therefore, we conclude that equilibrium informativeness when information is verifiable is weakly decreasing in ρ. Finally, consider the extreme case, $\rho = 1$. It's immediate to check that in this case, irrespective of weather information is verifiable or not, equilibrium informativeness coincides and it is equal to $\left(\frac{q-\mu_0}{1-\mu_0}\right)^{\frac{1}{2}}$. $\qquad \square$

A.3.6 Proof of Proposition 3

Let information be unverifiable. Fix $q' > q > \mu_0$ and consider $\rho \geq \frac{q'-\mu_0}{q'(1-\mu_0)}$. We want to show that the informativeness of truth-leaning equilibria under q' is higher than under q. To see this, note that ρ is large enough so that equilibria are strictly informative, for both q' and q. In particular, thanks to Proposition A3 and Proposition 2, we know that under q the equilibrium informativeness is equal to $\left(\frac{q-\mu_0}{1-\mu_0}\right)^{\frac{1}{2}}$ and, since $\frac{q-\mu_0}{1-\mu_0} < \frac{q-\mu_0}{1-\mu_0}$, we conclude that the informativeness of truth-leaning equilibria under q' is higher than under q. Now let information be verifiable. In this case, both $\bar{\rho}$ and $\left(\frac{q-\mu_0(\rho+q(1-\rho))}{(1-\mu_0)(\rho+q(1-\rho))}\right)^{\frac{1}{2}}$ are increasing in q. Therefore, irrespective of ρ, equilibrium informativeness under q' is higher than under q. $\qquad\square$

B Design

Figures B17 and B18 show the relevant screenshots from our experiment. The top panel of Figure B17 shows the sender's decision screens. Figure B18 shows the receiver's decision screens. The receiver could see the exact probability of each message by hovering the mouse cursor over the communication plan. The bottom panel of Figure B18 shows the Feedback screen. All relevant information were reported to both players, with the exception of the sender's choices in the Revision stage.

Figure B16: Screenshots from U100S: Commitment and Gussing Stage

Communication Stage

Here you choose your COMMUNICATION PLAN.
After you click Confirm, we will communicate the plan you chose to the Receiver.

If the ball is RED:

Send Message	with probability:
Red	52 %
Blue	24 %
No Message	24 %

0   25   50   75   100

If the ball is BLUE:

Send Message	with probability:
Red	17 %
Blue	28 %
No Message	55 %

0   25   50   75   100

CONFIRM

Update Stage

Here you can Update your COMMUNICATION PLAN.
The Receiver cannot see how you UPDATE your COMMUNICATION PLAN.

The Ball is Red.

The message that you will send will be generated:

- With Probability 80%, from the COMMUNICATION PLAN you chose at the previous stage.
- With Probability 20%, from the UPDATE you choose now.

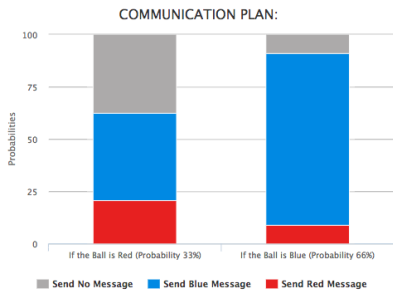| Send Message	with probability:
Red	37 %
Blue	40 %
No Message	23 %

0   25   50   75   100

CONFIRM

Figure B17: Screens 1 and 2, Treatment U80

Guessing Stage

The message you will receive will come:

- with probability 20%, from the UPDATE, that you can't see.
- with probability 80%, from the COMMUNICATION PLAN you see below:

COMMUNICATION PLAN:

Choose your GUESSING PLAN:

If I Receive Message...   ...my guess will be:

The Ball is Red   RED   BLUE

The Ball is Blue   RED   BLUE

No Message   RED   BLUE

Summary:

Ball Color	Message Sent	Origin	Guess	Your Payoff	Opponent's Payoff
xxx	xxx	xxx	xxx	xx Dollars	xx Dollars

You selected this COMMUNICATION PLAN:
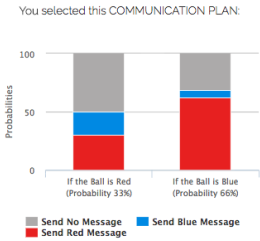
the Receiver selected this GUESSING PLAN:

If I receive Message Red, I will guess "xxx"
If I receive Message Blue, I will guess "xxx"
If I receive No Message, I will guess "xxx"

When you are done, press Continue to proceed:

CONTINUE

Figure B18: Screens 3 and 4, Treatment U80

16

C   Instructions for V80

In this section, we reproduce instruction for one of our treatment, V80. These instruction were read out aloud so that everybody could hear. A copy of these instructions was handout to the subject and available at any point during the experiment. Finally, while reading Section C.2.1, screenshots similar to those in Appendix B, were shown to subjects, to ease the exposition and the understanding of the tasks.

C.1   Welcome:

You are about to participate in a session on decision-making, and you will be paid for your participation with cash vouchers (privately) at the end of the session. What you earn depends partly on your decisions, partly on the decisions of others, and partly on chance. On top of what you will earn during the session, you will receive an additional $10 as show-up fee.

Please turn off phones and tablets now. The entire session will take place through computers. All interaction among you will take place through computers. Please do not talk or in any way try to communicate with other participants during the session. We will start with a brief instruction period. During the instruction period you will be given a description of the main features of the session. If you have any questions during this period, raise your hand and your question will be answered privately.

C.2   Instructions:

You will play for 25 matches in either of two roles: **sender** or **receiver**. At the beginning of every Match one ball is drawn at random from an urn with three balls. Two balls are BLUE and one is RED. The receiver earns $2 if she guesses the right color of the ball. The sender's payoff only depends on the receiver's guess. She earns $2 only if the receiver guesses RED. Specifically, payoffs are determined illustrated in Table C9.

The sender learns the color of the ball. The receiver does not. The sender can send a message to the receiver. The messages that the sender can choose among are reported in Table C10.

Each Match is divided in three stages: Communication, Update and Guessing.

1. Communication Stage: before knowing the true color of the ball, the sender chooses a COMMUNICATION PLAN to send a message to the receiver.

17

	If Ball is Red		If Ball is Blue	
If Receiver guesses Red	Receiver $2	Sender $2	Receiver $0	Sender $2
If Receiver guesses Blue	Receiver $0	Sender $0	Receiver $2	Sender $0

Table C9: Payoffs

If Ball is Red:

- Message: *"The Ball is Red."*

- No Message.

If Ball is Blue:

- Message: *"The Ball is Blue."*

- No Message.

Table C10: Messages

2. Update Stage: A ball is drawn from the urn. The computer reveals its color to the sender. The sender can now UPDATE the plan she previously chose.

3. Guessing Stage: The actual message received by the receiver may come from the Communication stage or the Update stage. Specifically, with probability 80% the message comes from the Communication Stage and with probability 20% it comes from the Update Stage. The receiver will not be informed what stage the message comes from. The receiver can see the COMMUNICATION PLAN, but she cannot see the UPDATE. Given this information, the receiver has to guess the color of the ball.

At the end of a Match, subjects are randomly matched into new pairs. We now describe what happens in each one of these stages and what each screen looks like:

C.2.1 Communication Stage: (Only the sender plays)

In this stage, the sender doesn't yet know the true color of the ball. However, she instructs the computer on what message to send once the ball is drawn. In the left panel, the sender decides what message to send if the Ball is Red. In the right panel, she decides what message to send if the Ball is Blue. We call this a COMMUNICATION PLAN.

Every time you see this screen, pointers in each slider will appear in a different random initial position. The position you see now is completely random. If I had to

reproduce the screen once again I would get a different initial position. By sliding these pointers, the sender can color the bar in different ways and change the probabilities with which each message will be sent. The implied probabilities of your current choice can be read in the table above the sliders.

When clicking Confirm, the COMMUNICATION PLAN is submitted and immediately reported to the receiver.

C.2.2 Update Stage: (Only the sender plays)

In this Stage, the sender learns the true color of the ball. She can now update the COMMUNICATION PLAN she selected at the previous stage. We call this decision UPDATE. The receiver will not be informed whether at this stage the sender updated her COMMUNICATION PLAN.

C.2.3 Guessing Stage. (Only the receiver plays)

While the sender is in Update Stage, the receiver will have to guess the color of the ball. On the left, she can see the COMMUNICATION PLAN that the sender selected in the Communication Stage. By hovering on the bars, she can read the probabilities the sender chose in the Communication Stage. Notice that the receiver cannot see whether and how the sender updated her COMMUNICATION PLAN in the Update Stage. On the right, the receiver needs to express her best guess for each possible message she could receive. We call this A GUESSING PLAN. Notice that once you click on these buttons, you won't be able to change your choice. Every click is final.

C.2.4 How is a message generated?

With 80% probability	With 20% probability
The message is sent according to COMMUNICATION PLAN	The message is sent according to UPDATE
(Remember: COMMUNICATION PLAN is always seen by the Receiver)	(Remember: UPDATE is never seen by the Receiver)

C.3 Practice Rounds:

Before the beginning of the experiment, you will play 2 Practice rounds. These rounds are meant for you to familiarize yourselves with the screens and tasks of both roles. You will be both the sender and the receiver at the same time. All the

choices that you make in the Practice Rounds are unpaid. They do not affect the actual experiment.

C.4 Final Summary:

Before we start, let me remind you that.

- – The receiver wins $2 if she guesses the right color of the ball.

- – The sender wins $2 if the receiver says the ball is Red, regardless of its true color.

- – There are three balls in the urn: two are Blue (66.6% probability), one is Red (33.3% probability). After the Practice rounds, you will play in a given role for the rest of the experiment.

- – The message the receiver sees is sent with probability 80% using COMMUNICATION PLAN and with probability 20% using UPDATE.

- – The choice in the Communication Stage is communicated to the receiver. The choice in the Update stage is not.

- – At the end of each Match you are randomly paired with a new player.

D   Additional Material

Table D11: P-Values of Statistical Tests

| Model Subject Session Bootstrap	Linear RE Cluster	Linear RE RE	Pr(T)obit RE Cluster	Pr(T)obit RE RE	Linear FE Cluster CATs	Linear FE Cluster
Test						
$\Pr\left(red\|\mu < \frac{1}{2}\right) = \Pr\left(red\|\mu \geq \frac{1}{2}\right)$	0.000	0.000	0.000	0.000	0.011	0.012
$\Pr\left(red\|m = r, \mu < \frac{1}{2}\right) = \Pr\left(red\|m = r, \mu \geq \frac{1}{2}\right)$	0.000	0.000	0.000	0.000	0.010	0.014
$\Pr\left(red\|m = b, \mu < \frac{1}{2}\right) = \Pr\left(red\|m = b, \mu \geq \frac{1}{2}\right)$	0.010	0.000	0.003	0.000	0.047	0.078
Left panel Figure 10, all bars = 0 when ball is R	0.000	0.000				
Left panel Figure 10, all bars = 0 when ball is B	0.000	0.000				
$\phi_C^B = \phi_R^B$ in U80	0.000	0.000	0.000	0.996		
$\phi_C^B = \phi_R^B$ in V80	0.000	0.000	0.000	0.006		
Right panel Figure 10, r message bar = 0 when ball is R	0.000	0.000				
$\Pr\left(red\|m = r, \mu < \frac{1}{2}\right) = \Pr\left(red\|m = r, \mu \geq \frac{1}{2}\right)$ in U20	0.053	0.002	0.083	0.004	0.150	0.126
$\Pr\left(red\|m = r, \mu < \frac{1}{2}\right) = \Pr\left(red\|m = r, \mu \geq \frac{1}{2}\right)$ in U100	0.000	0.000	0.024	0.000	0.040	0.021
$\Pr\left(red\|m = r, \mu < \frac{1}{2}, U20\right) = \Pr\left(red\|m = r, \mu < \frac{1}{2}, U100\right)$	0.627	0.535	0.718	0.610		
$\Pr\left(red\|m = r, \mu \geq \frac{1}{2}, U20\right) = \Pr\left(red\|m = r, \mu \geq \frac{1}{2}, U100\right)$	0.000	0.001	0.002	0.003		
$\Pr\left(red\|m = n, \mu < \frac{1}{2}\right) = \Pr\left(red\|m = n, \mu \geq \frac{1}{2}\right)$ in V20	0.038	0.002	0.133	0.006	0.257	0.196
$\Pr\left(red\|m = n, \mu < \frac{1}{2}\right) = \Pr\left(red\|m = n, \mu \geq \frac{1}{2}\right)$ in V100	0.000	0.000	0.000	0.000	0.022	0.014
$\Pr\left(red\|m = r, \mu < \frac{1}{2}, V20\right) = \Pr\left(red\|m = r, \mu < \frac{1}{2}, V100\right)$	0.566	0.674	0.536	0.452		
$\Pr\left(red\|m = r, \mu \geq \frac{1}{2}, V20\right) = \Pr\left(red\|m = r, \mu \geq \frac{1}{2}, V100\right)$	0.000	0.000	0.000	0.000		
$\phi(V20) = \phi(V80)$	0.217	0.215				
$\phi(V80) = \phi(V100)$	0.001	0.020	0.258	0.451		
$\phi(U20) = \phi(U80)$	0.002	0.001				
$\phi(U80) = \phi(U100)$	0.696	0.676	0.486	0.441		
$\phi(V20) = \phi(U20)$	0.000	0.000				
$\phi(V80) = \phi(U80)$	0.000	0.000				
$\phi(V100) = \phi(U100)$	0.000	0.000	0.000	0.000		
$\phi^B(V20) = \phi^B(V80)$	0.156	0.130				
$\phi^B(V80) = \phi^B(V100)$	0.032	0.052	0.608	0.648		
$\phi^B(U20) = \phi^B(U80)$	0.000	0.000				
$\phi^B(U80) = \phi^B(U100)$	0.957	0.925	0.711	0.661		
$\phi^B(V20) = \phi^B(U20)$	0.000	0.000				
$\phi^B(V80) = \phi^B(U80)$	0.000	0.000				
$\phi^B(V100) = \phi^B(U100)$	0.000	0.000	0.000	0.000		
$\phi^B(U100) = \phi^B(U100H)$	0.144	0.116	0.205	0.180		
$\phi^B(U100) = \phi^B(U100H)$ in last 3 matches	0.052	0.038	0.061	0.056		
$\Pr\left(red\|\mu < \frac{1}{2}, U100\right) = \Pr\left(red\|\mu < \frac{1}{2}, U100H\right)$	0.069	0.053	0.026	0.026		
$\Pr\left(red\|\frac{1}{2} \leq \mu < \frac{3}{4}, U100\right) = \Pr\left(red\|\frac{1}{2} \leq \mu < \frac{3}{4}, U100H\right)$	0.008	0.110	0.011	0.125		
$\Pr\left(red\|\mu \geq \frac{3}{4}, U100\right) = \Pr\left(red\|\mu \geq \frac{3}{4}, U100H\right)$	0.001	0.014	0.008	0.045		

The p-values reported in the text are obtained by regressing the variable of interest on the relevant regressor (sometimes an indicator variable) with subject level random effects and clustering of the variance-covariance matrix at the session level. This specification has the advantage of being uniform (the same throughout the paper), it directly accounts for heterogeneity across subjects via the random effects (as the paper documents, there is clear evidence of heterogeneity between subjects), and it permits unmodeled dependencies between observations from the same session (see Fréchette (2012) where such possibilities are discussed). However, it does not directly account for the fact that we are many times dealing with a limited dependent variable. Also, clustering with a small number of clusters can lead to insufficient corrections (Cameron and Miller (2015) for a survey). But this

relies mostly on simulations that do not necessarily mirror the situation of most laboratory experiment. In particular, the extent of the problem is found to depend on the size of the within session correlation (see for example Carter et al. (2017)). For many experiment, such correlation can be expected to be low (once the appropriate factors are controlled for). Hence, we are more concerned with controlling for the source of dependencies across the observations of a given subject than for the within session correlations (see also Appendix A.4 of Embrey et al. (2017) for a discussion of these issues).

In Table D11 we document the robustness of the tests reported in the text by exploring alternative specifications. These include directly accounting for the limited nature of the dependent variable by using a probit or tobit when appropriate. When possible we also report bootsraped estimates that have been shown to perform better when the number of clusters is small (cluster-adjusted t-statistics or CAT) and allow for subject specific fixed-effects (Ibragimov and Müller, 2010). When we report those we also include results from a standard subject specific fixed-effects estimation with session clustering to provide a benchmark.

As can be seen, p-values are not systematically larger for CATs than with the "standard" clustering, nor are they very different when estimating a probit or tobit.[56] As a whole, results are fairly robust: out of the 35 hypothesis tested, for only six of them are results not the same for all tests reported (in the sense of being consistently significant–or not–at the 10% level). The few cases where there are differences are for the most part not difficult to make sense of. Two of them involve comparing V80 and V100, where the difference is small in magnitude. Hence, whether or not the difference is statistically significant is not clear, but the either way it is not large. In most other cases, the p-values are either under the 0.1 cutoff or slightly above.

---

[56]Note that if a tobit could have been estimated but is not reported, it means that the dependant variable was not actually censored.
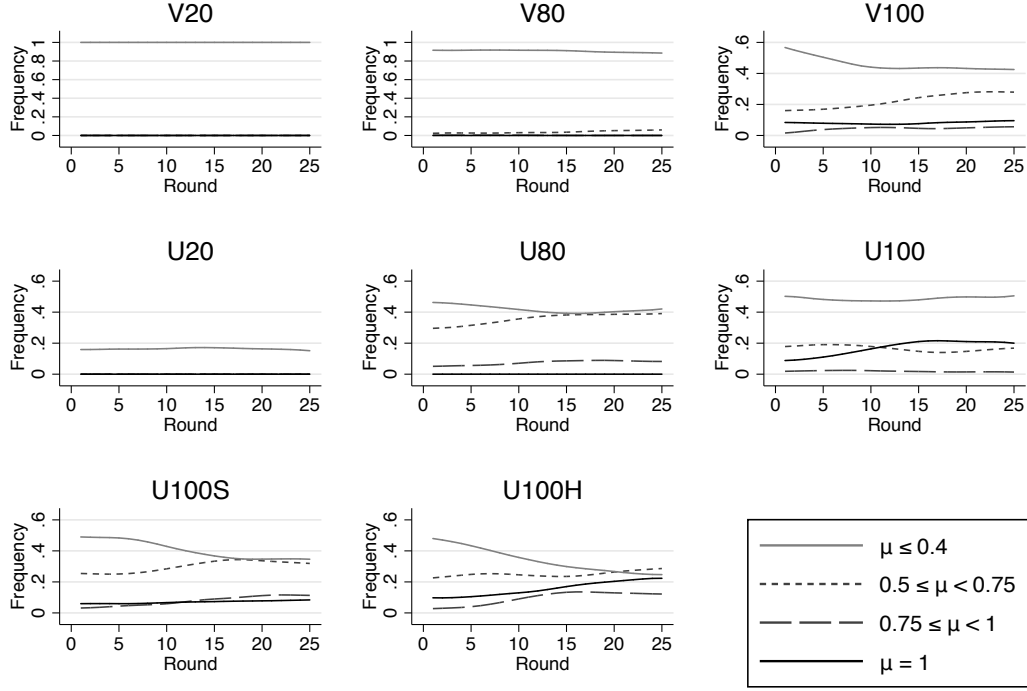
Posterior following a critical message: no message for V treatments and red message for U treatments
V20 and V80 are drawn with a different y-axis.

Figure D19: Frequency of Persuasive Messages Grouped by Posterior (μ)

The next two figures illustrate changes in behavior over the course of the experiment. Figure D19 does so for senders by coarsely separating sender strategies by the posterior they induce on red when sending a persuasive message; that is a n message under verifiable information and a textitr message under unverifiable information. Four message types are plotted: low information ($\mu < 0.4$), close to full-commitment equilibrium information ($0.5 \geq \mu < 0.75$), high information ($0.75 \geq \mu < 1$), and full disclosure ($\mu = 1$). The excluded category is close to, but below, full-commitment equilibrium information ($0.4 < \mu < 0.5$). As the figure shows, in some treatments there are very few changes over time (at least no change across these categories), for instance in treatment V20; while in others there are substantial developments over the course of the experiment. One such example is treatment U100H where senders move away from low information strategies toward more informative ones.
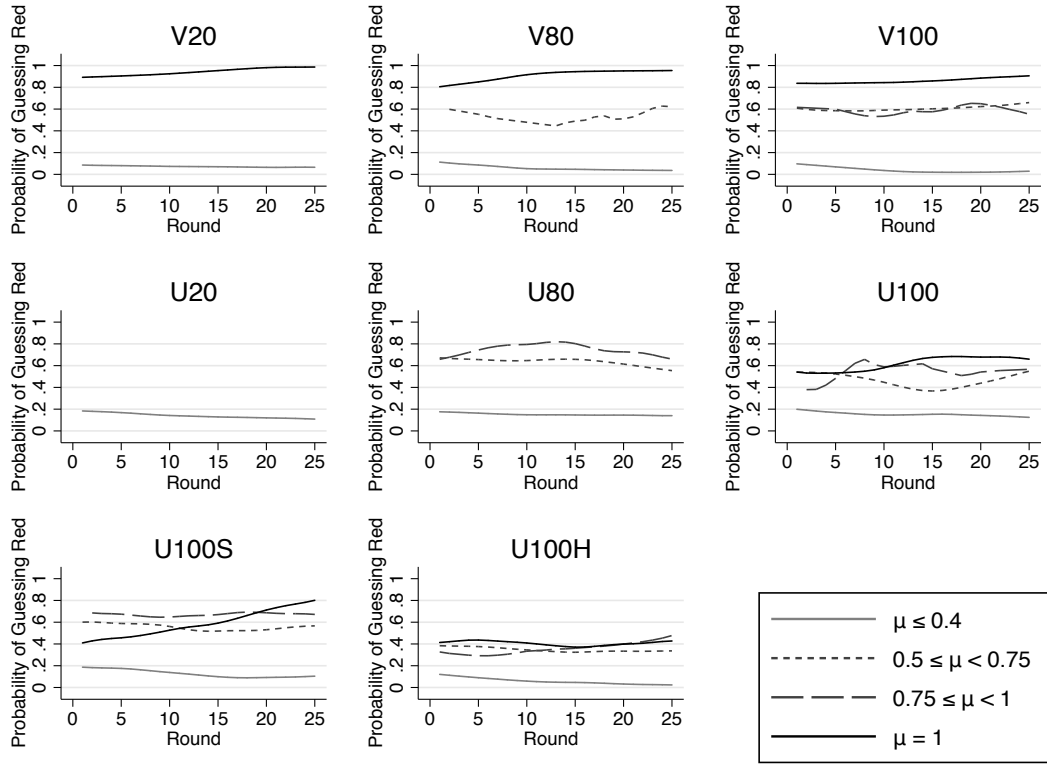
23

Figure D20: Frequency of Guessing *red* Grouped by Posterior (μ)

On the responder side, Figure D20 also displays changes in terms of the likelihood a given posterior leads to a guess of *red*. In all verifiable treatments, there is a slight increase in the probability of guessing *red* over rounds. At the other end, theres seems to be a generalized decrease over rounds of guessing *red* when the posterior is low.

Blue indicates a subject who's choices are best explained by the rule always guess blue.
Values are jittered slightly to make multiple overlapping thresholds distinguishable.
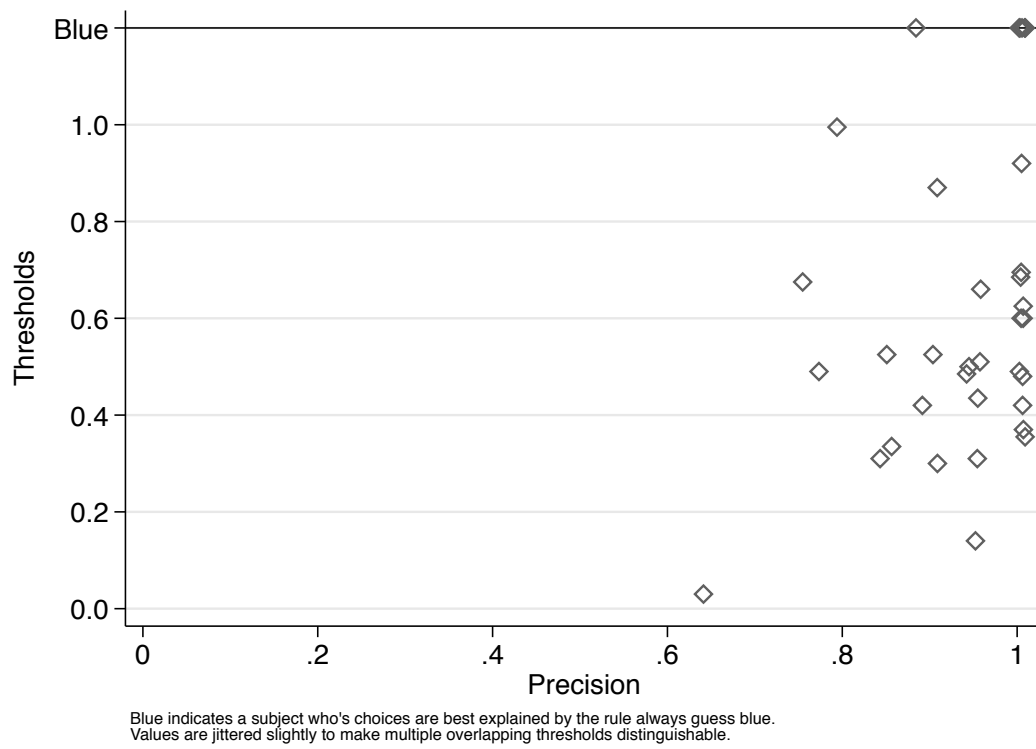
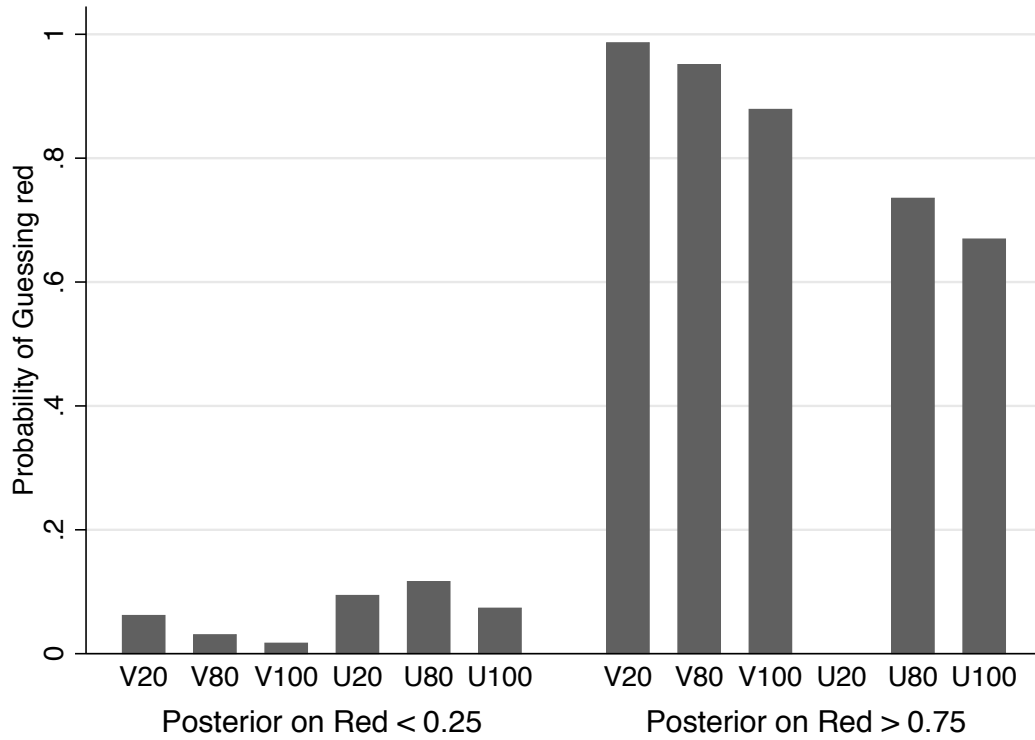Figure D21: Estimated Threshold and Precision for Treatment U100S

Figure D22: Probability of Guessing Red as a Function of Posterior

Voting patterns in all treatments are similar in that they are increasing in the posterior on red. They do display some revealing differences however. As can be seen in Figure D22, treatments with verifiable messages lead to more "certainty" in voting. When the posterior on red is low, the probability of guessing *red* is even lower in the verifiable treatments (it is already very low in the unverifiable treatments) and when the posterior is high, the probability is much closer to one in the treatments with verifiable messages.

Blue indicates a subject who's choices are best explained by the rule always guess blue.
Values are jittered slightly to make multiple overlapping thresholds distinguishable.
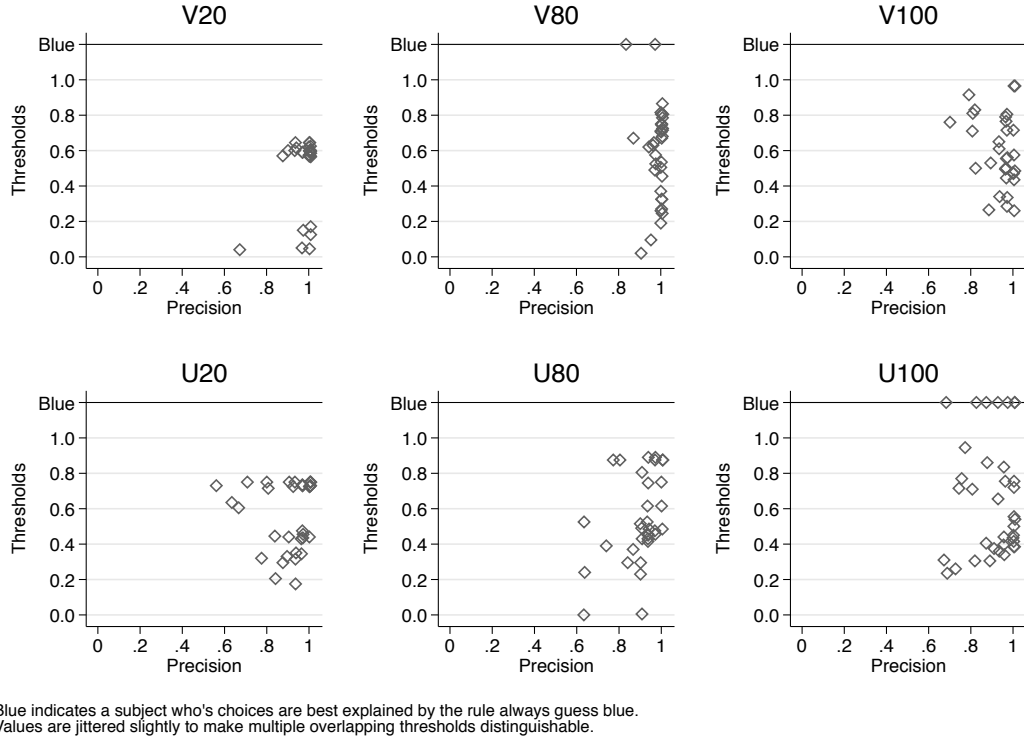
Figure D23: Estimated Threshold and Precision

Figure D23 illustrates the best fitting thresholds and their precision for the general treatments. Unlike for the U100S treatment, these are based on 30 choices per subjects (thus having a high precision is a more demanding test). Nonetheless, precision is still high, with the treatment with lowest precision still having 81% of subjects with 80% precision and across all treatments 90% of subjects meeting that criterion. The figure also shows that precision is particularly high when messages are verifiable. Indeed, under verifiable messages, 55% of subjects always choose in a way that is consistent with a threshold. That number is 24% for the treatments with unverifiable messages. The figure also confirms the finding of heterogeneity across receivers.

27

Black for subjects harder to convince than a Bayesian, gray for subjects easier to convince than a Bayesian.
Blue indicates a subject who's choices are best explained by the rule always guess blue.
Values are jittered slightly to make multiple overlapping thresholds distinguishable.
Only includes responders for whom the best fitting threshold accounts for at least 80% of choices.
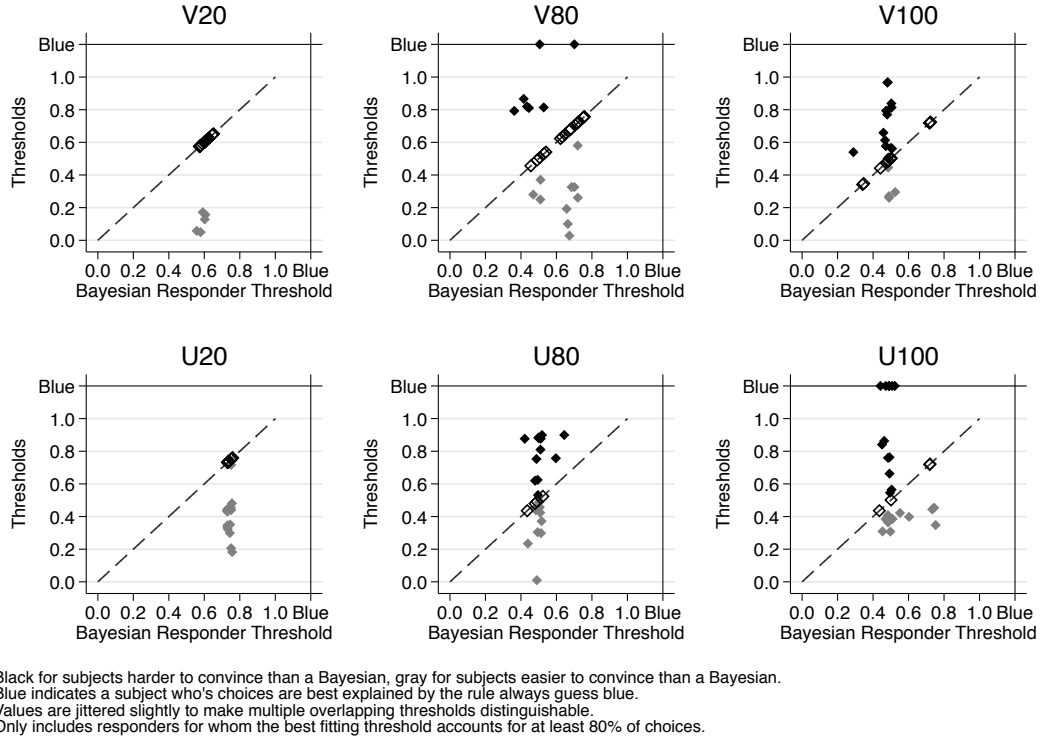
Figure D24: Estimated Threshold: Actual Receivers Against Bayesian

Figure D24 compares the estimated thresholds for subjects with good precision to what we would recover if the subjects were Bayesian. This confirms what was established in Section 2.2, namely that a non-trivial fraction of subjects are close to the behavior Bayesian receivers would exhibit, but there are also subjects who need a higher, and others lower, posterior to guess *red*. Note also that in our treatment that comes closest to the setup of cheap talk experiments, all deviations from Bayesian behavior indicate receivers who are gullible.

Prior experiments on communication (of the kind considered here) have mainly considered cheap talk and disclosure (see our Introduction for a list of references). Typical results involve: (1) Some transmission of information under cheap talk, although far from complete. This comes about both via (2) senders conveying more information than predicted and (3) receivers reacting to messages. (4) Less than full information transmission in disclosure environments. This is because (5) of a partial failure of unravelling. Our results are consistent with these earlier observations.

Correlations for treatments with $\rho = 0.2$ reported in Table 4 are in line with points 1 and 4: there is some information transmission in U20 (correlation $= 0.09$), and there is less than full information transmission in V20 (correlation $= 0.83$).
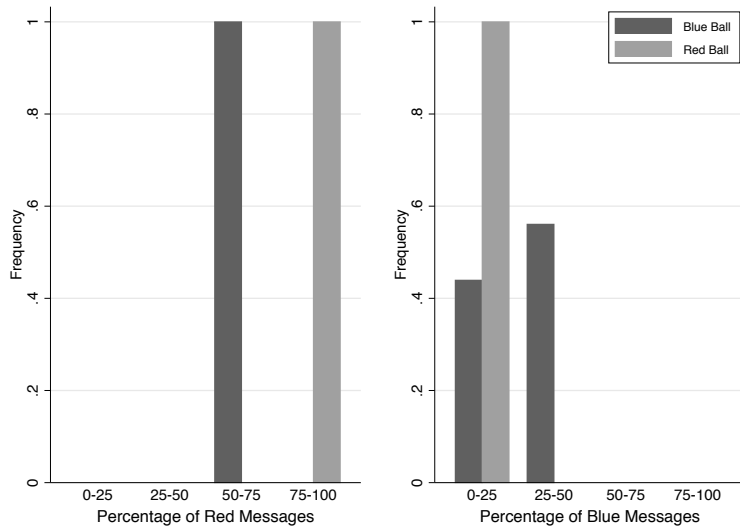

Figure D25: U20 Revision Stage

The parallel to point 2 can be evaluated in Figure D25. The figure shows that all strategies send a r message when the ball is *Blue* more than 50% of the time. However, they do not send a r message 100% of the time when the ball is *Blue*. In other words, all strategies misrepresent the state the majority of the time, but they also indicate the truth a fraction of the time.

Consistent with point 3, receivers in U20 are 29 percentage points more likely to guess *red* following a r message ($p-value < 0.01$) than a *blue* message. In other words, some receivers take messages at face value. This effect of message color is also found in other treatments in the case where both r and b messages should both mean that it is very likely the ball is *Red*. The right panel of Figure D26 considers such situations. Indeed in the U100 treatment, the effect of a r message that induces a posterior of more than 0.75 on red generate a 45 percentage points higher chance of guessing red than a similar b message. We also note that, although less pronounced, this phenomenon is nonetheless present in the simpler U100S treatment.
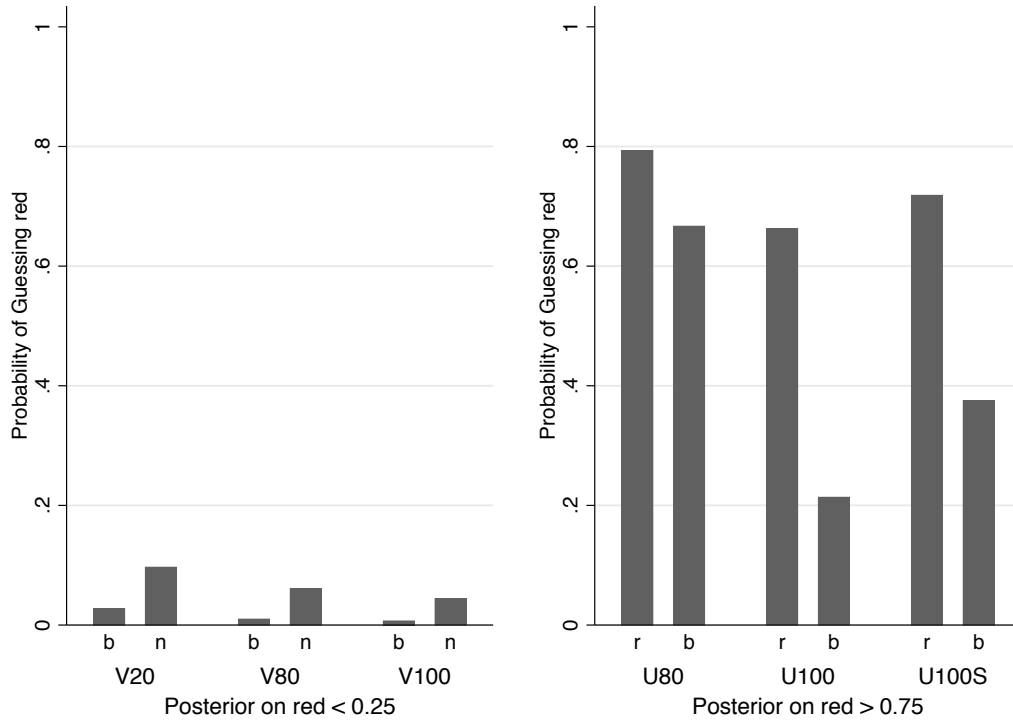
Figure D26: Probability of Guessing *red* as a Function of the Message

Similarly, in line with point 5, receivers in V20 are 7 percentage points more likely to guess *red* when receiving no message than when they received a blue message ($p-value < 0.05$). This effect is also found when commitment is available. As can be seen in the left panel of Figure D26, the probability of guessing *red* is higher after a *r* message in all three treatments (restricting attention to cases with equally low posteriors on red). In our environment, the effect of ball color in the unverifiable treatments is greater than the effect of a failure of unravelling in the verifiable treatments.
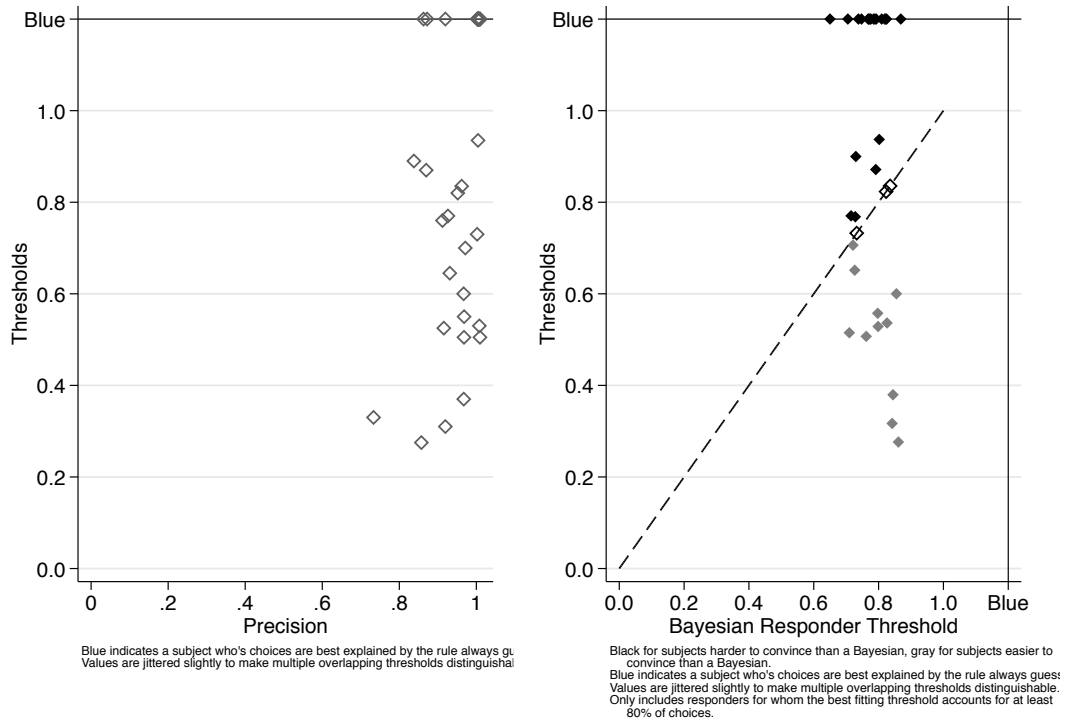
Figure D27: Estimated Threshold and Precision: U100H

Precision of best fitting threshold strategies in the U100H is very high: 97% of receivers with 80% precision and 47% with 100% precision. However, in this case, it is partly due to the fact that more receivers (as compared to other treatments) always guess *blue*. These are illustrated in the left panel of Figure D27.

Among subjects for whom the best threshold does not suggest always picking *blue*, the pattern is similar to other treatments. The right panel of Figure D27 shows there is heterogeneity in terms of how thresholds compare to what Bayesian receivers would do.
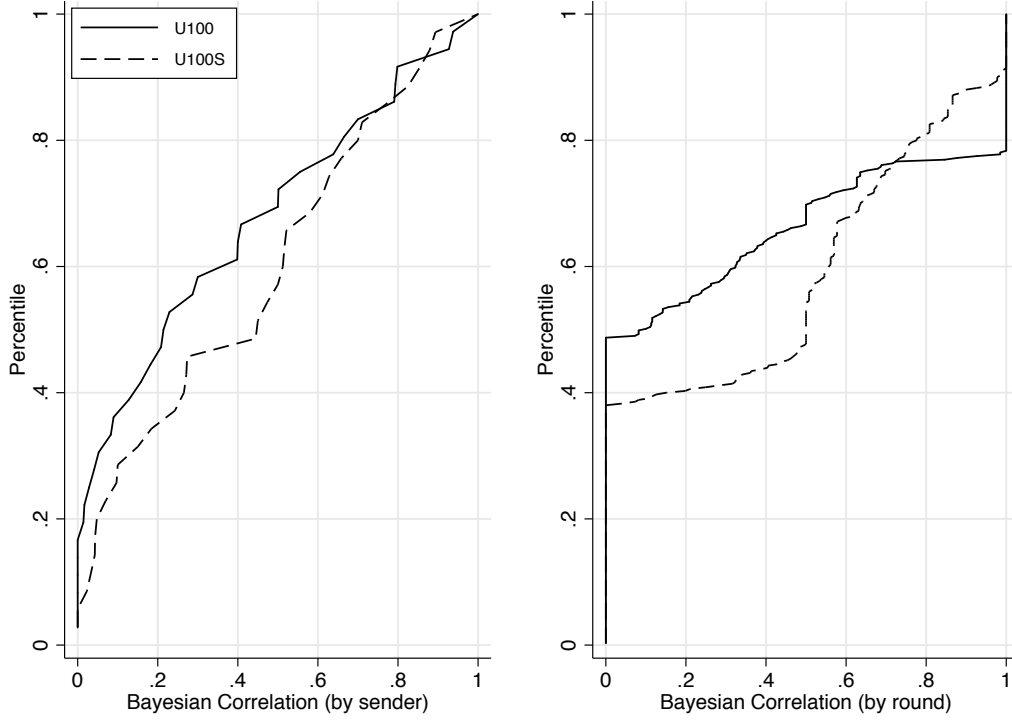
Figure D28: CDF of Bayes Correlation (ϕ^B): U100 and U100S

Figure D28 shows that behavior in the U100S treatment is similar to that in the U100 treatment. It also suggests slightly more information transmission in the U100S treatment (mean of subject averages is 0.41 in that treatment versus 0.33 in the U100 treatment). However, disaggregating the data further reveals one additional way in which U100S is closer to the theory than U100. The right panel of Figure D28 reproduces the CDFs of ϕ^B without first averaging at the subject level. This shows that under U100S fewer messages generate no correlation or full information. In addition, there is a higher density of messages that create exactly a correlation of 0.5. All of these differences make sender behavior in U100S closer to the theory than it is in U100.
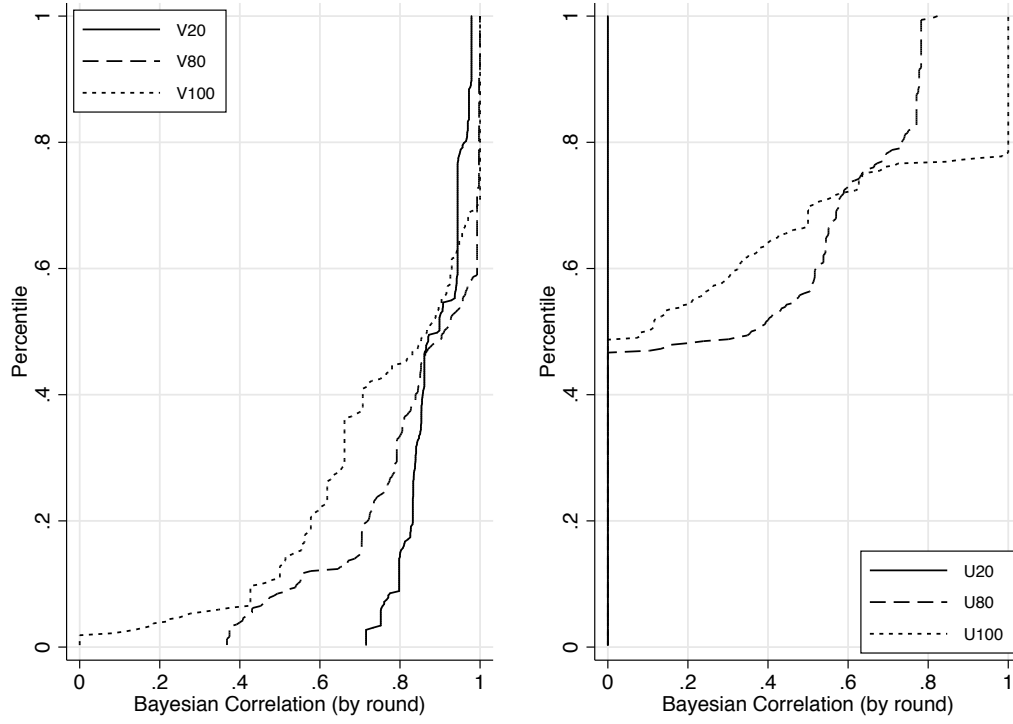
Figure D29: CDF of Bayes Correlation (ϕ^B)

Similarly, to the case above, not averaging correlations by subject produces different CDFs in other treatments as well. Nonetheless, the overall patern of cross treatments comparative statics is unchanged as can be seen in Figure D29.
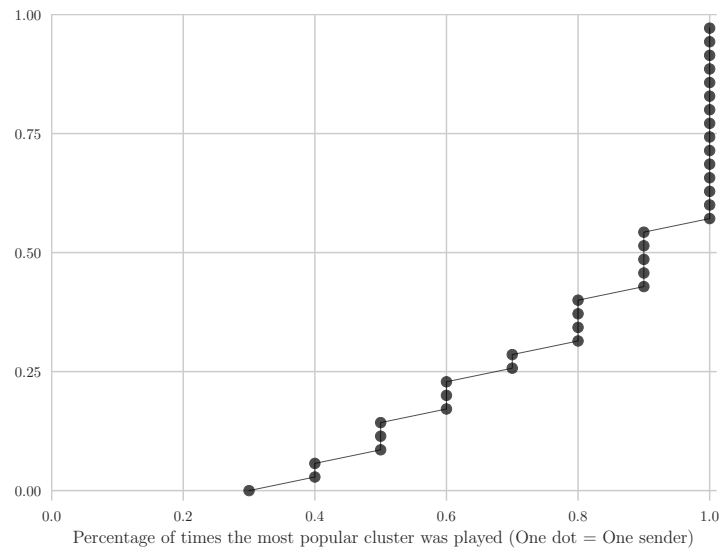
Figure D30: Persistence in senders' type

D.1 Thresholds

The data is composed of pairs of posteriors μ and guesses a for each receiver. We look for a threshold t that minimizes $\mathbb{1}\{a \neq \mathbb{1}\{\mu \geq t\}\}$ where a takes value 1 for *red* and 0 for *blue*. In other words, we find the treshold that rationalizes the greatest number of choices a subject has made.[57] We refer to the fraction of choices properly accounted for by the threshold as the precision. Given that the sample is finite and thresholds exists on the unit interval, there will be an infinite number of thresholds with the same precision. For instance, imagine a sample composed of two choices: a receiver that guessed *red* given a posterior of 0.57 and guessed *blue* when the posterior was 0.46. In that case, any threshold greater than 0.46 and less than or equal to 0.57 has a precision of 1. The figures report the average of the lowest and greatest threshold with the highest precision.

The theory assumes Bayesian receivers, i.e. agents who guess *red* for all posteriors of 0.5 or higher. However, even if our subjects were perfect Bayesians, we are unlikely to estimate their thresholds to be 0.5 due to the finite nature of the sample. For instance, in the example highlighted above is consistent with a Bayesian receiver, yet the estimated threshold would have been 0.515. Hence, when comparing the receivers in our experiment to the Bayesian benchmark, we do this by computing what threshold we would have estimated given the sample of posteriors if the receiver was perfectly Bayesian.

D.2 Variance of Induced Posteriors

In the paper, we used ϕ^B, namely the correlation between the state and the guess of Bayesian receiver, to measure the informativeness of a sender's strategy. In Section 2, we discussed the merits of this measure and how it relates to existing literature. In this appendix, we reevaluate our main comparative-static exercise from Section 4 with an alternative measure of informativeness, the variance of induced posteriors. Of course, this measure is highly correlated with ϕ^B, but it differs in that it does not require the specification of a payoff function for the receiver. In fact, given an information structure π, one can compute the induced distribution over posterior beliefs $\tau \in \Delta(\Delta(\Theta))$. The variance of τ, namely $\psi := \mathbb{E}_\tau\big((\mu - \mu_0)^2\big)$, is what we call the variance of induced posteriors. This measure can be seen as an alternative measure for the informativeness of π.

In Table D12, we report the average posterior variance across treatments together with the theoretical predictions. As for Table 4, this measure of informativeness moves in the direction predicted by the theory. Namely, it increases in

---

[57]This is akin to a perceptron in machine learning, see for instance Abu-Mostafa et al. (2012).

Table D12: Average Posterior Variance per Treatment

$\psi^\star$ – Theoretical Predictions	Commitment (ρ)				ψ – Empirical Posterior Variance	Commitment (ρ)		
	20%	80%	100%			20%	80%	100%
Verifiable	0.22	0.08	0.05		**Verifiable**	0.18	0.17	0.15
Unverifiable	0.00	0.05	0.05		**Unverifiable**	0.02	0.05	0.06

treatments with unverifiable information and it decreases in treatments with verifiable information. Yet, as for ϕ^b, the point-predictions are far from the empirical averages. In particular, senders in $V100$ appear to be overly informative relative to the prediction and there is a large gap between $V100$ and $U100$. In Figure D31, we report the CDF of sender-average ψ. These results are in line with those in Figure 10.
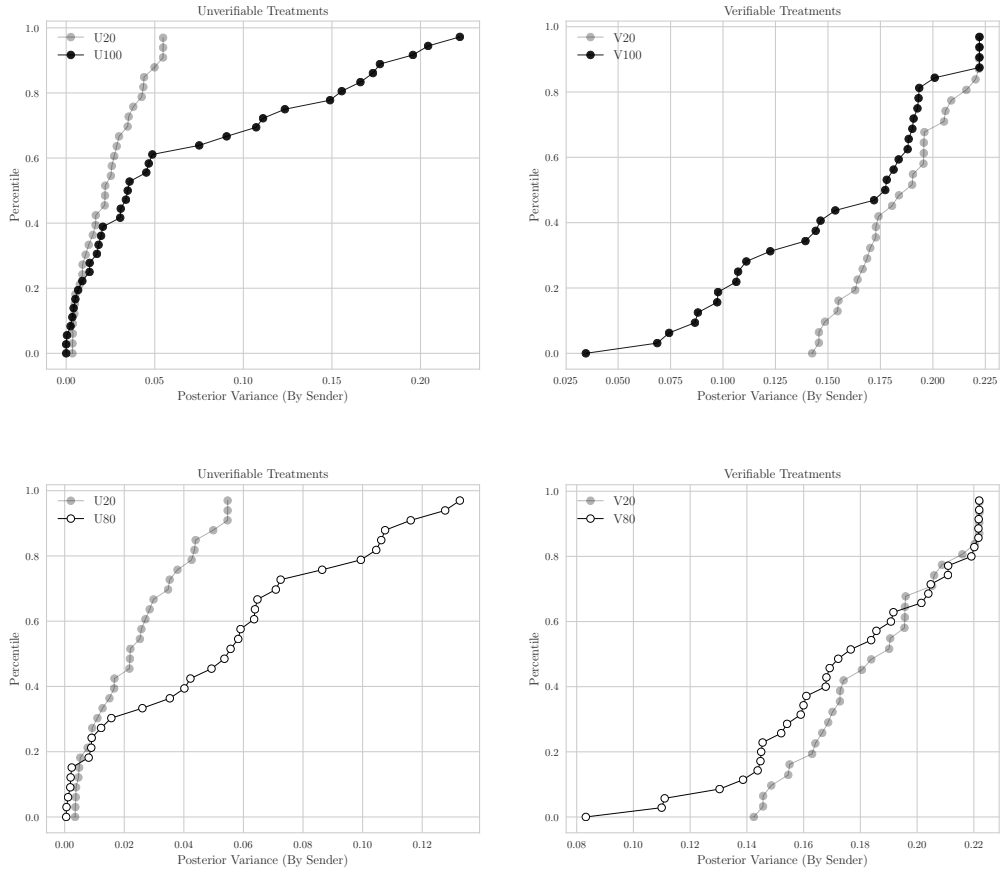

Figure D31: CDF of Sender-Average Variance of Induced Posteriors (ψ)

D.3 Quantal Response Equilibrium – Robustness

In this section, we explore the robustness of the QRE estimates reported in Table 6. Our methodology to estimate (λ_S, λ_R) relies on the discretization of the senders' strategy space to obtain a finite set of sender's strategies Π_k. Clearly, this is common practice when estimating QRE on games with a continuum of strategies (see, for instance, Camerer et al. (2016)). In our game, the discretization of the sender's strategy space is particularly challenging because it is multi-dimensional and data are asymmetrically distributed in this space (e.g. see Figure 6 for a simple example). For these reasons, choosing a uniform grid leads to computational problems, e.g. clusters with no data points. To cope with this, we take a more structured approach and discretize the strategy space via a k-means clustering algorithm. This algorithm finds k clusters and k representative strategies within each cluster that minimize the sum of the distances between each observed strategy and the representative strategy of the cluster to which it belongs.[58] The discretization of the strategy space leaves one main degree of freedom, namely the number of clusters k. In Table D13 we report estimates of (λ_S, λ_R) for various choices of k. Overall, we note that our estimates tend to stabilize as k grows large. Moreover, the ordinal rankings in λ_i across treatments are quite stable. For example, λ_S is consistently smaller in U100 relative to all other treatments; U100H is the treatment where senders are best-responding more effectively; λ_R^{U100} is consistently smaller than λ_R^{V100}; both λ_S and λ_R are higher in $U100S$ than in $U100$, capturing the fact that $U100S$ is a simplified version of $U100$.

So far, we computed the set Π_k on a treatment-by-treatment basis. That is, the set Π_k is computed by feeding data of that treatment only into the k-mean clustering algorithm. An alternative approach is to compute the set Π_k by clustering senders' strategies *across* treatments. In this way, we compute a single Π_k for all treatments, further improving our ability to compare QRE estimates across treatment. The results of this exercises are shown in Table D14. From the table, we note that, while QRE estimates change, the ordinal rankings are similar to those of 6. For example, λ_S is consistently lowest in U100 and highest in U100H; λ_R is consistently lowest in U100H and highest in V100.

---

[58]Computationally, the k-means clustering algorithm is initialized with a random draw of the set of representative strategies Π_k. This set is then iteratively updated and refined until convergence. In some instances, this initial randomness can affect final clusters. To eliminate this spurious dependence, we estimate (λ_S, λ_R) a 100 times and report the averages.

Table D13: QRE λ Estimates

	λ_S	λ_R	λ_S	λ_R	λ_S	λ_R	λ_S	λ_R	λ_S	λ_R
	$k=8$		$k=9$		$k=10$		$k=11$		$k=12$	
U100	1.23	1.36	0.11	1.36	0.01	1.34	0.05	1.33	0.18	1.32
V100	1.85	1.77	2.19	1.78	2.27	1.77	2.33	1.77	2.16	1.78
U100S	2.57	1.46	2.54	1.49	2.98	1.52	2.92	1.53	2.32	1.54
U100H	1.75	1.28	2.06	1.28	2.53	1.27	2.86	1.26	2.63	1.27
	$k=13$		$k=14$		$k=15$		$k=16$		$k=17$	
U100	0.39	1.31	0.44	1.3	0.4	1.3	0.36	1.29	0.32	1.3
V100	2.04	1.78	2.0	1.79	1.67	1.78	1.67	1.79	1.59	1.79
U100S	2.06	1.54	1.81	1.54	1.56	1.54	1.51	1.55	1.39	1.55
U100H	2.89	1.26	2.85	1.26	2.69	1.25	2.65	1.25	2.68	1.24
	$k=18$		$k=19$		$k=20$		$k=21$		$k=22$	
U100	0.32	1.3	0.33	1.3	0.34	1.31	0.35	1.3	0.36	1.31
V100	1.57	1.78	1.7	1.79	1.84	1.79	2.02	1.79	1.99	1.79
U100S	1.38	1.54	1.61	1.54	1.85	1.54	2.06	1.54	2.11	1.54
U100H	2.55	1.24	2.43	1.24	2.36	1.23	2.5	1.23	2.34	1.23

Table D14: QRE λ Estimates

	λ_S	λ_R	λ_S	λ_R	λ_S	λ_R	λ_S	λ_R	λ_S	λ_R
	$k=8$		$k=9$		$k=10$		$k=11$		$k=12$	
U100	0.28	1.38	0.05	1.42	0.0	1.4	0.0	1.38	0.0	1.35
V100	0.0	1.99	0.74	1.97	0.93	1.83	1.89	1.83	1.96	1.84
U100S	3.11	1.43	3.46	1.41	3.08	1.44	2.88	1.48	2.63	1.49
U100H	4.35	1.21	4.58	1.24	4.48	1.24	4.46	1.26	4.95	1.22
	$k=13$		$k=14$		$k=15$		$k=16$		$k=17$	
U100	0.0	1.35	0.0	1.34	0.0	1.34	0.02	1.33	0.03	1.33
V100	1.44	1.84	1.96	1.82	1.73	1.8	2.34	1.81	2.3	1.81
U100S	2.59	1.51	2.65	1.53	2.75	1.54	2.97	1.54	3.06	1.54
U100H	4.49	1.22	4.33	1.22	4.28	1.22	4.48	1.22	4.26	1.22
	$k=18$		$k=19$		$k=20$		$k=21$		$k=22$	
U100	0.04	1.33	0.14	1.33	0.13	1.33	0.18	1.32	0.15	1.32
V100	2.48	1.81	1.47	1.8	1.1	1.81	1.03	1.81	1.11	1.81
U100S	3.0	1.53	2.74	1.53	2.7	1.54	2.53	1.53	2.57	1.53
U100H	4.19	1.22	3.91	1.23	3.87	1.23	3.77	1.23	3.71	1.23