Rules and Commitment in Communication

Guillaume R. Fréchette
NYU

Alessandro Lizzeri
NYU

Jacopo Perego
Columbia U

October 9, 2018

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 or not information is verifiable; and the senders *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 or not. 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 rational behind the role of commitment and, more specifically, for the Bayesian persuasion model of Kamenica & 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.

We thank Andreas Blume, Santiago Oliveros, Salvatore Nunnari for useful comments. Fréchette and Lizzeri gratefully acknowledge financial support from the National Science Foundation via grant SES-1558857.

1 Introduction

The goal of this paper is to study how *rules* and *commitment* affect the amount of information that can be transmitted between a sender and a receiver who have conflicting interests. The structure of our experimental design allows us to jointly analyze models that share an underlying structure regarding preferences and information, but that are distinguished either by the rules governing communication, i.e., whether the sender can lie about the content of her information, or by the extent to which the sender can commit to her communication plan. The basic structure we propose generates a rich set of qualitative predictions regarding the communication strategies that are used by senders, the way the receiver is supposed to react to these strategies, as well as the amount of information that is revealed in equilibrium. With minimal differences between treatments, our common structure ranges from models of cheap talk (Crawford & Sobel (1982)), to models of disclosure (Grossman (1981), Milgrom (1981), Jovanovic (1982), Okuno-Fujiwara et al. (1990)), to models of Bayesian persuasion (Kamenica & 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.

In order to set the stage, we begin with our first contribution: the experimental analysis of a Bayesian persuasion game (Kamenica & Gentzkow (2011)). 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 induce a posterior conditional on the high message that makes the receiver indifferent between choosing the low and the high action, thereby maximizing the ex ante probability that the receiver chooses the high action. Broadly speaking, the experimental results show that on average subjects behave in ways consistent with the logic of the model. However, in the experiment, subjects can be classified into three approximately equal-sized clusters. One cluster exhibits equilibrium-like behavior in that senders choose strategies that are close to the equilibrium predictions. Another cluster displays behavior that results in drastic over-communication relative to the equilibrium benchmark. The last cluster exhibits behavior inducing important under communication. We will discuss below other interesting aspects of behavior in this treatment. The main conclusions we draw from this treatment are

that the model provides a useful framework to think of the important forces at play in such environment, but behavior is also very heterogeneous, and that it can be classified into meaningful and persistent clusters. However, from this treatment it is difficult to draw conclusions about the degree to which agents understand commitment and whether they manage to exploit their commitment power. This leads us to propose a general structure to generate a rich set of qualitative predictions that allows us to draw more meaningful conclusions about the effects of commitment.

Our second contribution is to introduce *partial* commitment. 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. This is referred to as the *commitment stage*. After observing the state, with probability $(1-\rho)$ the sender is given the opportunity to secretly revise her plan of action. This is referred to as the *revision stage*. The higher is ρ, the higher is the probability that the sender will not be able to revise her strategy, hence the higher the extent to which she is committed to her initial plan of actions. In the limit case in which $\rho = 1$ the sender has full commitment and outcomes converge to the Bayesian persuasion model (Kamenica & Gentzkow (2011)) studied in our first treatment. The third contribution is to consider verifiable and unverifiable communication in the same framework. 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.[1] The first scenario corresponds to models of cheap talk (e.g., Crawford & Sobel (1982)), while the latter corresponds to models of disclosure with verifiable information (e.g., Grossman (1981), Milgrom (1981), Jovanovic (1982), Okuno-Fujiwara et al. (1990)).[2]

Novel comparative statics arise from this setup. We focus our attention on the effect that the degree of commitment has on the amount of information that is revealed in equilibrium and how this depends on the verifiability of messages. While in any specific environment and application it is hard to know (and measure) the exact extent of commitment available to an agent, it is natural to think that this can vary and may depend on observable correlates such as the protocols and the

[1]The sender misreports her private information when she sends messages that are *false*. A message is false if none of the statements it contains is true. E.g., provided that the ball is blue, message "the ball is red or black" is false. We will formalize this in Section 3.1.

[2]The (un)verifiability of communications could be modeled in more detail. For instance, verifiable messages could be the equilibrium result of the sender's aversion to lying: when lying is cheap, communication is unverifiable; when the cost of lying is sufficiently high, communication becomes verifiable. Alternatively, verifiability could be modeled as access, with some probability, to an authority that verifies the truthfulness of the report. When that probability is zero, communication is unverifiable, and when it is one, communication is verifiable.

frequency of communication. Thus, it seems important to study how communication varies with the extent of commitment.

When messages are unverifiable, the sender seeks commitment because it allows her to credibly tell the truth, at least some of the time. That is, commitment *increases* the amount of information that is revealed in equilibrium. In contrast, in verifiable environments, i.e. when any informative message needs to correspond to the state, the sender still seeks commitment, but for the opposite reason: to credibly hide information, at least some of the time. That is, commitment *decreases* the amount of information that is revealed in equilibrium. We also show that, when senders have full commitment, verifiability has no impact on the amount of information that is communicated. That is, equilibrium informativeness is exactly the same regardless of the verifiability of messages. However, the senders implement this outcome with very different strategies. Thus, the model generates a rich set of predictions that we bring to the laboratory.

Our main results are the following. We first show that subjects understand the power of commitment: senders figure out how to exploit commitment and receivers how to react to it. We show that senders understand commitment by contrasting their behavior in the commitment stage to their behavior in the revision stage in a treatment with partial commitment, and we exploit the contrasting predictions with verifiable versus unverifiable information. The theory predicts that in the case of unverifiable information the sender 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 supports this prediction of the theory. In addition, the receiver should understand that information conveyed in the commitment stage is more meaningful when the level of commitment is higher, and this is what we find in the data: in higher commitment treatments receivers are more responsive to information from the commitment stage, and, again, this ranking is reversed when messages are verifiable. We then consider how commitment affects overall equilibrium informativeness by comparing different levels of commitment. We find that subjects behave in ways that are consistent with theory, that is, informativeness decreases with commitment in the verifiable treatments and increases with commitment in the unverifiable treatments. However, quantitatively there are significant departures from the theory: informativeness does not rise enough in the unverifiable treatments and it does not decrease enough in the verifiable treatments. Furthermore, commitment seems to work better for the sender when communication is unverifiable. 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, the equilibrium amount of communication is the same in the two treatments. We also find that, in the unverifiable treatments with a substantial amount of commitment, receivers are excessively skeptical. This is partly in contrast

with prior literature on cheap talk (see the review by Blume et al. (2017)). From a policy perspective, this is a novel justification for making it harder for senders to misreport their information. Finally, we consider an orthogonal prediction of the theory regarding the consequences of changing the receivers' payoffs so as to change what we call the persuasion threshold, i.e., the minimal amount of information required to persuade the receivers to choose the action desired by the sender. The theory predicts that the sender conveys more information when the persuasion threshold is higher, and this 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, it is crucial to make sure that all sources of variations coming from seemingly unimportant details of the design are reduced to a minimum, so that differences in outcomes in the data can be imputed to differences in the treatments. In order to do this, 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 it 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 we do not investigate the relationship between the informativeness of communication and the degree of preference alignment between the sender and the receiver. Rather than preference alignment, we focus on the effect of commitment on the amount of information that is revealed in equilibrium.[3]

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 introduced in Kamenica & Gentzkow (2011). This model has become influential in the recent theoretical literature, e.g. Gentzkow & Kamenica (2014), Alonso & Camara (2016), Gilligan & Krehbiel (2016), etc. Evaluating how the degree of commitment affects outcomes is one way to experimentally evaluate the model of Bayesian persuasion.

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 e.g., Gilligan & Krehbiel (1989), Gilligan & Krehbiel (1987); and the production of evidence to a jury (Kamenica & Gentzkow (2011), Alonso & Camara (2016)). Models of disclosure of ver-

[3]In doing this, we are also addressing recent theoretical contributions on persuasion under partial commitment, such as Min (2017).

ifiable information have been used to study the disclosure of quality by a privately informed seller, for instance, via warranties,[4] of the contents of financial statements by a firm,[5] and in many other contexts. Dranove & Jin (2010) survey the literature on product quality and the disclosure of information.

There are a number of experimental papers on 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 information 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. In the experiment, in contrast, 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 & Wang (2006) find that Senders are overly truthful and they also find 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: consistently with the theory, they find that 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]

Conversely, experiments on the disclosure of verifiable information typically find that there is 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. This in turn leads senders to underprovide information, thereby undermining the unraveling argument.[7] There are also some papers that 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 & Leslie (2003) study the consequences of mandatory hygiene grade cards in restaurants. They show that hygiene

[4]E.g., Grossman (1981).

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

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

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

cards lead to increases in hygiene scores, that demand becomes more responsive to hygiene, and to lower food borne illness hospitalizations.

Our paper is one of three new experimental investigation of Kamenica & Gentzkow (2011). Nguyen (2017) and Au & 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 & 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 where preferences are such that agents have other-regarding concerns. And 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, it is far from certain that a receiver will guess in a way that benefits the sender (in both studies around 50%). These results are also consistent with our findings, which suggests that these results are robust given that all three implementations are fairly different.

2 A Benchmark Treatment of Bayesian Persuasion

It is useful to start our analysis from our simplest treatment. By doing so, we will uncover a number of basic facts characterizing how senders and receivers behave. Importantly, the main themes highlighted in this section also hold in the general framework of Section 3.

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, that

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 *information 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.[8]

Table 1: Payoffs

	State (θ)			
Guess (a)	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 one; conditional on state $\theta = B$, she commits to sending messages r and b with equal probability.[9] 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 one half following a message r. Thus, the receiver guesses *blue* following a message b and is willing to guess *red* following message r since she is indifferent between *red* and *blue*.[10] Two simple features of equilibrium stand out. First, the sender benefits from commitment. When based exclusively on her prior information, 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.

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

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

[10]In equilibrium, the receiver must choose *red* with probability one following message r since otherwise the sender would choose an information structure that induces a slightly higher posterior conditional on message r, but then there would be no best response for the sender.

Implementation in the laboratory.[11] At the beginning of each session, instructions were read aloud, subjects were assigned a fixed role (sender or receiver). In each session subjects play 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 to 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 show-up fee, were \$34 (ranging from \$14 to \$52) per session. Figures 1 and 2 are examples of the screens from the two main stages.

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 1. These probabilities are updated in real-time in the cells above the sliders. The receiver observes the information structure chosen by the sender (as in Figure 2) and makes a guess for each possible message (strategy method). The specific probabilities of 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.[12]

2.2 Measuring Informativeness

The "amount" of information that the sender transfers to the receiver, namely the *informativeness* of her communication strategy, represents a variable of central interest in our analysis. As customary in the literature, we measure informativeness as the correlation coefficient between two random variables: the color of the ball and the receiver's guess.[13] We denote this variable by ϕ. To fix ideas, suppose that the sender transfers no information, then her final guess is independent of the message received and, therefore, the correlation between the state and the receiver's guess will be zero. If in contrast the sender truthfully disclosed the color of the ball to the receiver, then her final guess would be perfectly correlated with the state. However, unlike the rest of the literature, the nature of our game and its implementation al-

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

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

[13] See for instance Cai & Wang (2006), Forsythe et al. (1999), and Wang et al. (2010).

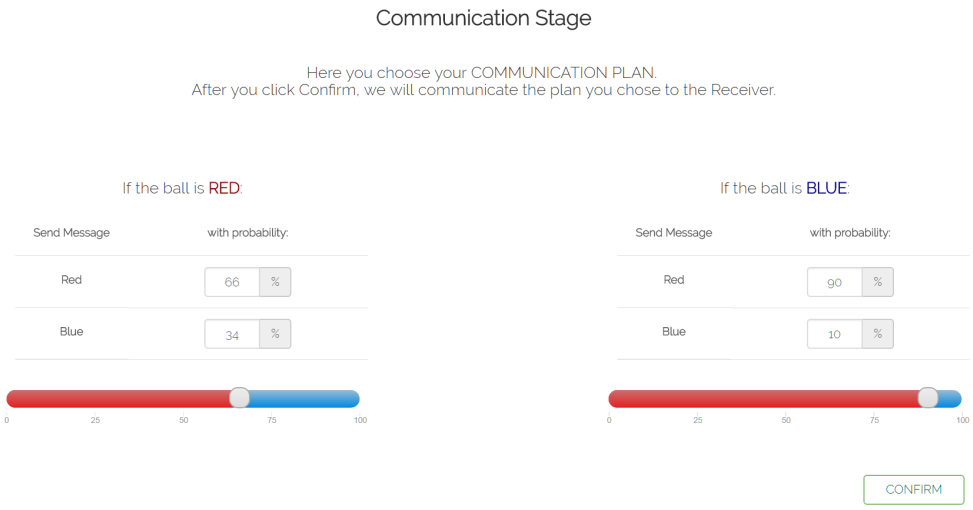


Figure 1: Design - Commitment Stage

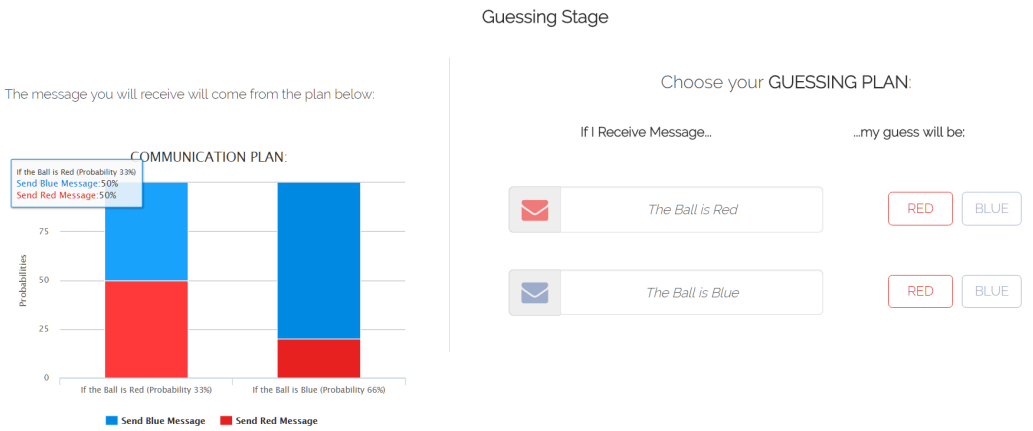


Figure 2: Design - Guessing Stage

lows us to leverage the power of the strategy method to obtain a significantly more precise measure of the correlation. Instead of using the *realized* state and guess, we use the observed strategies of senders and receivers to analytically compute the correlation coefficient (specifically the phi coefficient or mean square contingency coefficient). For the purpose of computing this correlation, it is as if we had an infinite sample of states and messages in each round.[14] Note that the correlation coefficient implied by the equilibrium strategies is 0.50.

This way of measuring informativeness has the potential drawback of combining the behavior (and the mistakes) of both senders and receivers. In particular, ϕ compounds the potential inability of the sender to communicate, with the potential unresponsiveness of the receiver to information. 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 sometimes use an alternative measure of informativeness, denoted ϕ^B, which is the correlation coefficient implied by the sender's strategy combined with the guesses of a hypothetical *Bayesian* receiver.

2.3 Main Results

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 sender behavior and then proceed with an analysis of receiver behavior.

2.3.1 Sender Behavior: Types and Informativeness

Describing sender behavior is challenging as the game is complex and strategies are high-dimensional objects. We approach the description of this behavior in several steps, starting at a more aggregate level, and moving to a less aggregated one. We begin by studying the informativeness of sender behavior.

Figure 3 plots the distribution of $\bar{\phi}_i^B$, the sender-specific average correlation coefficient with Bayesian receivers . The overall across-sender average correlation is 0.41 and the median is 0.45. Although this average is fairly close to the equilibrium value of 0.5, the distribution in Figure 3 clearly shows a high level of heterogeneity: some senders rarely reveal any information, some others consistently reveal almost all the information. However, there is also a non-negligible group of subjects that convey some, but not all, of the information.

One drawback of focusing on the correlation with Bayesian receivers is that it hides potentially useful information. For example, a sender who generates a poste-

[14] Simulations we have done suggests that the improvement in precision is non-trivial and that samples need to be large for the estimates of the Pearson correlation to stabilize.

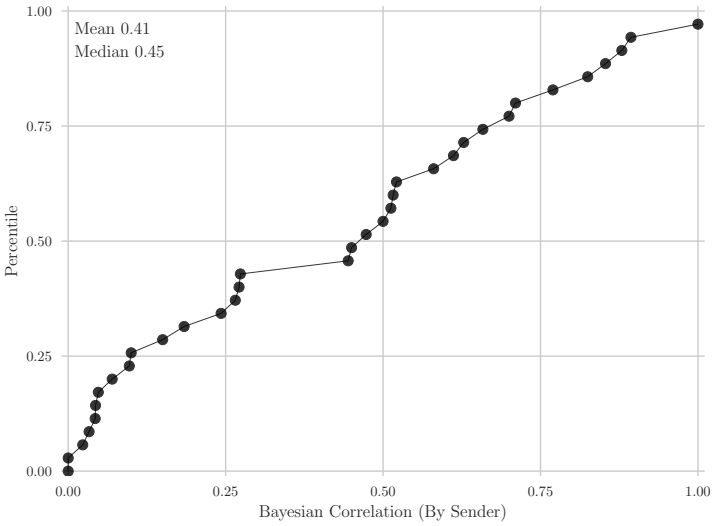


Figure 3: CDF of Subject Average Bayes Correlation ($\bar{\phi}_i^B$)

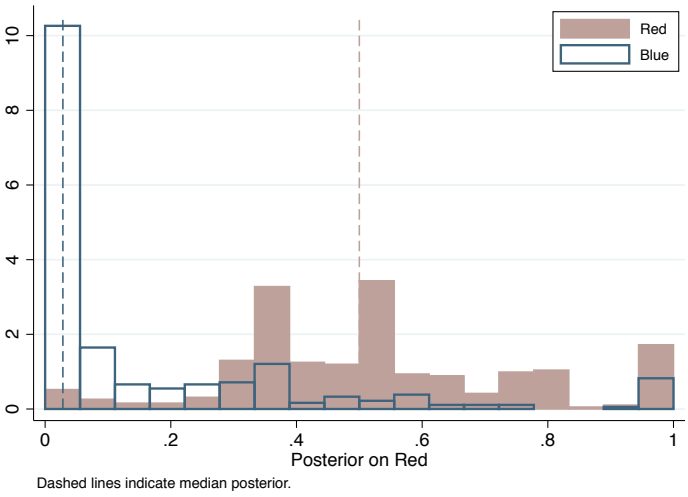


Figure 4: Histogram of Posterior on Red by Message

rior conditional on message r that is just below 0.5 does convey some information to the receiver. However, this posterior leads to a correlation of zero because the Bayesian receiver would choose *blue* in both states following such a posterior; i.e. the same correlation as if the sender had conveyed no information (a posterior of 1/3). We therefore report in Figure 4 the empirical distribution of Bayesian posteriors that are induced by the observed senders' strategies.[15] The figure reveals a few important facts. First, consistent with equilibrium prediction, 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. In contrast, the posterior conditional on a red message is highly dispersed. However, the most com-

[15]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, i.e. $\mu(m) := \mu_0(\theta = R|m)$.

mon posterior is close to 0.5, in line with the equilibrium prediction. The spikes in this distribution of posteriors at 1/3 and at 1 represent clusters of strategies that we discuss next.

There are many possible strategies that can generate a particular amount of informativeness. We now describe the strategies actually chosen by the senders. In our setting, strategies are highly dimensional objects; making it difficult to summarize them. We tackle this by grouping the data into groups of similar strategies. To do this we use a k-means clustering analysis of senders' probabilities of sending each message as a function of the state.[16] The results indicate that about 90% of the observed choices can be organized in three *clusters*, whose representative strategies, or *types*, are displayed in Figure 5.[17]

The three largest clusters explain, respectively, 23%, 35% and 24% of the sample of observed senders' strategies in this treatment. These strategies share a 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. This suggests that our clustering analysis captures a meaningful economic feature in senders' behavior. To see this, we compute the average correlation coefficient with Bayesian receivers, i.e. ϕ^B, 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). This means that the three clusters identify three substantially different *styles* of communication. The first one is particularly truthful and reveals a lot of information. The last one, instead, is uninformative. Finally, the intermediate (and most prevalent) cluster, 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.

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

[16] K-means clustering ((MacQueen 1967)) is a commonly used method to group data–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 to the closest center, and the centers are iterated on to minimize the total within cluster variance.

[17] The figure also displays a "residual cluster" that gathers all the remaining, harder to categorize, observations. In this exercise, we use clusters for descriptive purposes, and not for estimation. Therefore, determining the best number of clusters is not of particular importance. Nonetheless, we establish the "robustness" of the clusters we reported by producing them using two different starting values, and we obtain identical results for the two methods. One approach is initialized by using random groups, while the other uses the output of a clustering exercise on the Bayesian correlation (ϕ^B) as starting groups. The number of clusters is determined using the elbow method. However, an additional consideration is that starting with five clusters, the smallest cluster only has 11 observations (less than 4% of the data), making the results unreliable.

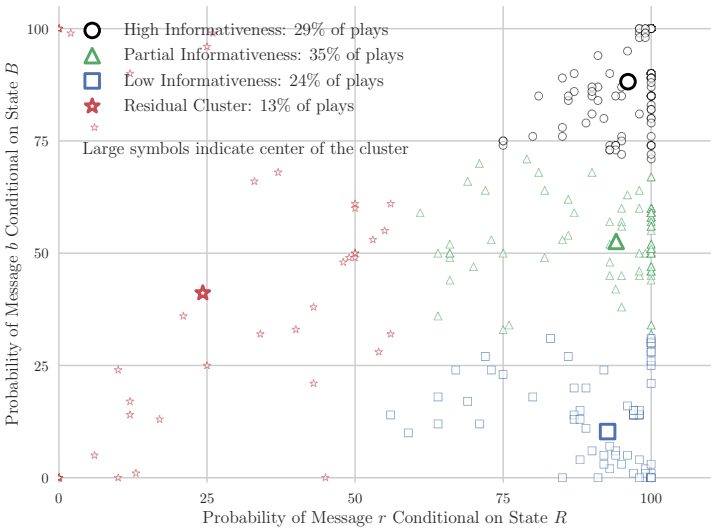


Figure 5: Sender's Strategies Grouped in Clusters

belongs to the same cluster nine times out of ten (more details in Figure D32 in Appendix D).

To summarize, despite the complexity of the decision that senders face, their behavior features strong regularities. Specifically, we have identified three prevalent and persistent types of senders in our data. They all predominantly use a "natural" language: that is, conditional on the state being R, they almost exclusively send message r. From the perspective of a Bayesian receiver, this implies that a message b is convincing evidence that the state is B. The most common sender strategy is of the equilibrium type. However, other important types of sender behavior are also observed with variations in how persuasive their message r is. In particular, for some it is highly persuasive while for others it is not at all. Next we describe receiver's behavior.

2.3.2 Receivers' Behavior

How do receivers respond to the information conveyed by senders' strategies? To answer this question, we start our analysis at the aggregate level. On average, receivers react to a higher posterior $\mu(m)$ by guessing *red* with higher frequency, as illustrated in Figure 6. That is, consistent with theoretical predictions, receivers are more persuaded to guess *red* by messages that carry more evidence in favor of the state being R. Specifically, for posteriors above $\frac{1}{2}$, receivers guess *red* 57% of the time, whereas they do so 11% of the time otherwise ($p \leq 0.01$).[18]

[18]Unless noted otherwise, all statistical results are established allowing for random-effects at the subject level and clustering 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 these suggest that session-effects are not important in this setting.

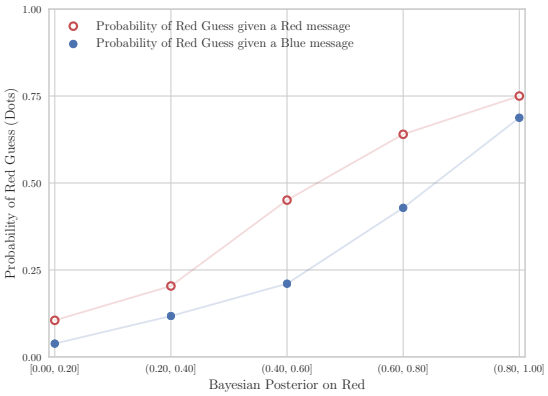


Figure 6: Probability of Guessing Red by Posterior and Message

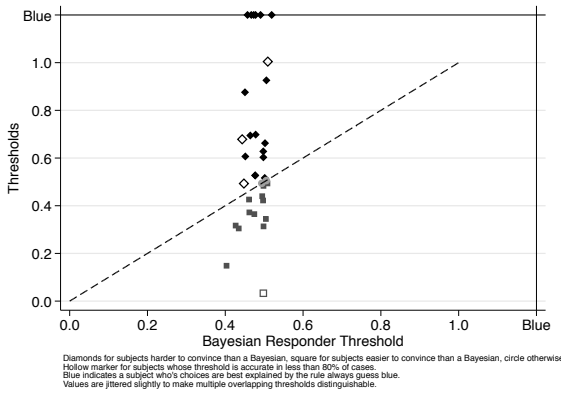


Figure 7: Estimated Thresholds: Actual Receivers vs Bayesians

This monotonicity is, however, a particularly mild requirement for receivers' rationality. Given the payoffs in our experiment, a Bayesian receiver should respond by guessing *red* to any posterior $\mu(m) \geq \frac{1}{2}$, and by guessing *blue* otherwise. Clearly, the aggregate evidence from Figure 6 does not fulfill this stronger requirement of rationality. One way to see this is to analyze the aggregate response conditional on the color of the message. When $\mu(m = r) \geq \frac{1}{2}$, receivers guess *red* 62% following a r message and 38% of the time following a b messgae. 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. Another way to see inconsistencies with the Bayesian paradigm is to note that, when provided with conclusive evidence that the state is R, i.e. even when $\mu(m)$ is arbitrarily close to 1, some receivers nonetheless guess *blue*, at least some of the time. To summarize, we can conclude that 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 for example Charness & 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.

While not being Bayesian, a receiver's behavior may exhibit systematic patterns in how it deals with the information received, as summarized by the posterior belief. In particular, receivers may follow a *threshold strategy*, guessing *red* if and only if the induced posterior is weakly above 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 suggests that she requires stronger evidence to guess that the ball is *red*. With our data, we can estimate, for each receiver, the threshold

that rationalizes the greatest fraction of her guesses.[19] We find that the behavior of many subjects is consistent with a threshold rule. Almost half of 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 7 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 was Bayesian.[20] As the figure shows, there is heterogeneity in receivers' behavior. Dots lying above the 45 degree line indicate receivers who are reluctant to guess *red*, despite the evidence. In contrast, the points below the 45 degree line indicate subjects who are credulous and too eager to guess *red*, despite the lack of evidence. The aggregation of this heterogeneous threshold behavior is partly responsible for the smoothness of aggregate responses to the posterior that is displayed in Figure 6. It is also notable that Figure 7 shows a sizable fraction of receivers that exhibits behavior consistent with the Bayesian benchmark: one quarter of the receivers have thresholds within five percentage points of exhibiting behavior consistent with a Bayesian receiver; the number increases to one third if we are more permissive and allow for a band of 10 percentage points around the Bayesian receiver.

Summing up, receivers' behavior is overall responsive to information (Figure 6). At an aggregate level, however, their responsiveness is not as pronounced as it would be for a Bayesian receiver. Figure 7 shows that in fact the vast majority of receivers behaves in ways that are consistent with threshold strategies, and does so in a persistent way. Importantly, there is heterogeneity in these thresholds, which explains the muted aggregate responsiveness. In particular, although many subjects are close to the Bayesian benchmark, others are too compliant while some too unwilling. Note that this observation that many subjects are more difficult to convince than theory suggests is in contrast to the findings from prior cheap talk experiments (Blume et al. 2017).

2.3.3 A Best-Response Analysis

The evidence on the receivers' behavior that we just documented raises a natural question: is the *equilibrium* strategy prescribed for senders close to a best-response

[19] 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 observation. 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.

[20] Notice in fact that given the finite nature of the data, even a Bayesian receiver could have an estimated 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 D23 in the appendix presents the estimated threshold and their respective precision.

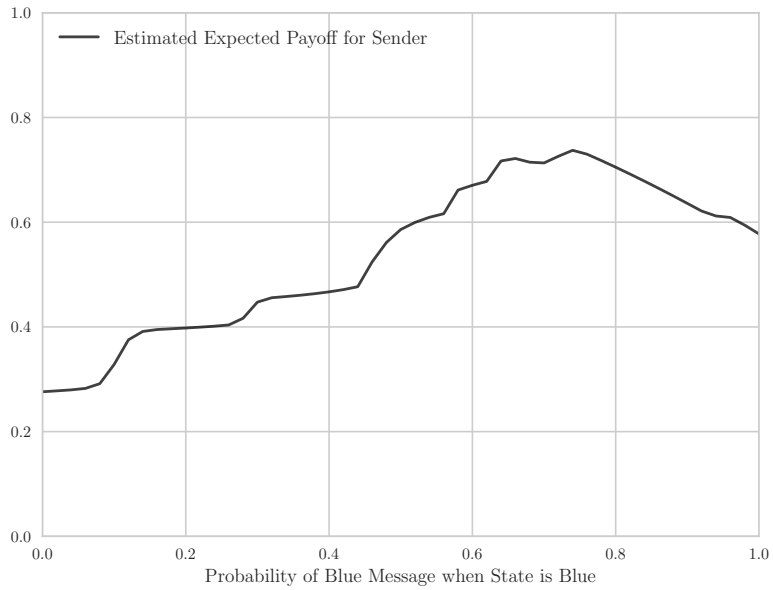


Figure 8: Expected Payoffs and Correlation

to the non-equilibrium behavior of our receivers? And if not what is the best-response? To answer these questions, we 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 a set 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 can accomodate the key strategies reported in Figure 5. That familly of strategies is the following: if $\theta = R$ the strategy will send message r with certainty. However, if $\theta = B$, we consider a continuum of different possibilities, one for each possible mixture between r and b. Therefore, we are able to parametrize each of these strategies with a number in the unit interval.[21] Figure 8 displays the expected payoff for these hypothetical senders' strategies, as a function of the level of mixing conditional on state B. Figure 8 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. It does not come as a surprise to see that the equilibrium strategy–which constitutes a knife-edge case in which the (Bayesian) receiver is indifferent–is not the best response to a population of human receivers. Instead, the best-response consists in overshooting a bit, that

[21]For example, when message b is sent with probability zero, the strategy is entirely uninformative, and all its induced posteriors are equal to the prior. Instead, when message b is sent with probability one, the strategy is perfectly informative. In this case, the induced posteriors are either 1, for the red message, or 0, for the blue message. Finally, the equilibrium strategy is a special case of this class of hypothetical strategies, as it involves a 50% mixture between r and b.

is, providing more information than required by the equilibrium benchmark. The figure also shows that receivers' departures from Bayesian behavior leads to a payoff function for senders that is flatter and smoother that if senders faced a population of Bayesian receivers.

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 takeaway which emerges from the analysis is that, despite a large heterogeneity in both the senders' and the receivers' behavior, the vast majority of subjects behave in meaningful and systematic ways. On the one hand, senders face a non-trivial task, namely the design of an information structure. Nonetheless, our analysis shows that most senders do engage in communication strategies that are sophisticated, and do so by employing a natural language. Senders differ among each other mainly in the exact amount of information that they are willing (or able) to transfer to the receivers. On the other hand, receivers do react to information and understand the basic trade offs that different information structures entail. Their behavior is also heterogeneous, but it is overall consistent with the use of threshold strategies. More specifically, receivers differ among each other in the amount of information they require in order to be successfully persuaded. 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. Instead, the sender should lie just enough to maximize her odds, while maintaining her credibility. Figure 8 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.

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 will describe the experimental design and the equilibrium outcomes that obtain under a specific parametrization of our model, the one that we bring to the laboratory. In the Appendix, we show that the predictions extend to a more general environment. The model that we present in Section 3.1 builds on the one we discussed before in two different ways. First, it weakens the commitment assumption by allowing for *partial* commitment. Second, it applies to settings with both *verifiable* and *unverifiable* information, two

different communication "rules" that have policy relevance.

The value of such a richer model is twofold. First, the interaction between partial commitment and communication rules generates particularly clear tests that we use to assess whether subjects understand commitment and how they take advantage of it. As will become clear later in this section, these tests are simply infeasible in a model with full commitment. In contrast, our richer framework enables us to go well beyond our analysis of Section 2.3, providing qualitative ways to evaluate the role of commitment in communication. Second, the more general framework presented in this section organizes under the same umbrella models of cheap talk, information disclosure, and Bayesian persuasion. These models, which form the backbone of the literature on strategic communication, assume either no commitment or full commitment. In any specific application, it is not always easy to understand whether commitment is a justified assumption. However, it seems reasonable to think that some settings are more amenable to commitment than others, for instance because participants are more likely to engage in repeated interactions. Thus, it is useful to understand the consequences of partial commitment and how communication varies with the degree of commitment.

3.1 Theory

Let $\Theta = \{\theta_L, \theta_H\}$ be the state space. There is a common prior μ_0 that denotes the probability that the state is θ_H. There are two players: a sender and a receiver. The sender has information, 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 game has three stages: the sender is active in the first two, while the receiver is active in the third. The receiver chooses actions in a binary set $A = \{a_L, a_H\}$. The receiver's 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. Given these parameters, a Bayesian receiver would choose action a_H whenever her posterior that the state is θ_H is larger than q. Thus, we call q the *persuasion threshold*.

We assume that the prior μ that the state is high is such that $\mu < q$ so that, absent any information, the receiver would choose a_L. Without this assumption the sender would have no reason to attempt to communicate with the receiver.

The sender's preferences are given by $v(a) := I(a = a_H)$, i.e., the sender receives a positive payoff only if the receiver chooses a_H. Thus, the sender would like the receiver to choose the high action.

The sender sends messages to the receiver from a subset of $M = \{\theta_L, \theta_H, n\}$.

The interpretation of these possible messages is that two of the messages represent statements about the states, while the last one is an "empty" message that has no relation to the state. In the experiment we call n "no message." Allowing for this message is important in the case of verifiable information, otherwise the sender would have no choice to make. Let $M^\theta \subseteq M$ be the set of messages that the sender can use in state θ. We say that information is *unverifiable* if $M^\theta = M$ for all θ. We say that information is *verifiable* if $M^\theta = \{\theta, n\}$ for all θ. Thus, when information is unverifiable, the sender is allowed to send any message, including any lie about the content of her information. When information is verifiable, the sender cannot lie about her type but can choose not to report the type by choosing the empty message n.

Messages are generated by the sender in two possible stages, a *commitment stage* in which she publicly chooses an information structure before observing the state, and a *revision stage* in which, after observing the state, she can secretly revise her strategy. More specifically, the game unfolds in the following way:

1. **Commitment Stage**. In this stage, the sender chooses an information structure, which is given by a commitment strategy: $\pi_C : \Theta \to \Delta(M^\theta)$ that defines the probability that the sender will send a specific message for any given realization of her information.

2. **Revision Stage**. The sender learns the state $\theta \in \Theta$. She now has the chance to revise her initial strategy π_C, by specifying a new probability distribution $\pi_R : \Theta \to \Delta(M^\theta)$.

3. **Guessing Stage**. The receiver chooses $a : M \times \Pi_c \to A$: the receiver observes the commitment strategy π_C but does not observe the revision strategy π_R. The receiver also observes a message m, which is generated with probability ρ from π_C and $(1-\rho)$ from π_R. This parameter ρ is exogenous and common knowledge. The receiver makes her best guess about θ, given each possible message $m \in M$ she might receive (and her knowledge of π_C and ρ).

It is useful to note that ρ is a measure of the sender's commitment: For high values of ρ, the choice made at the commitment stage is likely to be the relevant one, the one determining the final message. For low values of ρ, instead, the choice made at the revision stage is the one that is likely to matter.[22]

This game includes interesting classic models as special extreme cases. When $\rho = 0$ and information is unverifiable, this is a cheap talk model. In the binary model,

[22]Equivalently, one can think of the sender as having an *opportunity* to revise her commitment strategy, 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.

in this case, the unique equilibrium outcome involves no information transmission. When $\rho = 0$ and information is verifiable, this is a model of disclosure. The unique equilibrium in this case involves full information transmission. When $\rho = 1$ and information is unverifiable, this is a model of Bayesian persuasion. The unique equilibrium outcome involves partial information transmission.

A quantity of interest in our analysis is the amount of information that is transmitted from the sender to the receiver. We measure the amount of informativeness of communication in two ways: the correlation between the state and the action, and the dispersion of posteriors by state. In our discussion of the data these two measures are both useful and highlight different aspects of the phenomena of interest. However, our main theoretical results are independent of the specific measure of informativeness. When we say that there is a positive amount of information in equilibrium we mean that the equilibrium messages are correlated with the state. When we say that there is less than maximal information, we mean that this correlation is less than one.

Our first result considers a given level of commitment ρ and provides a contrast between the informativeness of communication at the commitment and at the revision stage.

Proposition 1. *(i) There is a $\widehat{\rho}$ such that, if $\rho > \widehat{\rho}$, in equilibrium, there is positive information in unverifiable treatments and less than maximal information in verifiable treatments. (ii) Consider ρ such that $\widehat{\rho} < \rho < 1$. Under verifiable information, more information is transmitted at the revision stage than in the commitment stage; under unverifiable information less information is transmitted at the revision stage than in the commitment stage.*

This result highlights the 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.

In order to understand this result it is useful to first consider the extreme cases: when $\rho = 0$, full disclosure takes place under verifiable information and no information is communicated under unverifiable information. When $\rho = 1$, it turns out that both scenarios lead to the same outcome of Bayesian persuasion resulting in partial information revelation (see Proposition below). The intuition for Proposition 1 is then the following. Under both verifiable and unverifiable information, the sender would like to commit to persuade the receiver to choose the high action as often as possible, and this requires partial information revelation. However, at the revision stage, the sender is unable to resist the temptation to use her opportunity to "cheat" and manipulate information in her favor. Under verifiable information this opportunity implies full information disclosure in the revision stage; under unverifiable information, it implies always stating the same favorable signal regardless of

the state. When ρ is high, some commitment is possible because the revision stage cannot completely undo the positive effect of the commitment stage, explaining part (i) of the result. However, the revision stage partially undoes the information communicated in the commitment stage, and this undoing goes in opposite directions in verifiable relative to unverifiable information, explaining part (ii) of this result. Part (ii) also says that the role of commitment in facilitating information transmission is completely reversed when we move from non verifiable to verifiable information.

Our next result describes how equilibrium informativeness changes with the degree of commitment, and how this depends on the verifiability of messages. For this and for the following result, when we compare the outcomes of different games, equilibrium multiplicity is a potential issue. As we show in the appendix, in the case of unverifiable information there is a unique equilibrium outcome. However, in the case of verifiable information, when ρ is sufficiently high but lower than one, there are multiple (similar) equilibrium outcomes. We focus on equilibria that are undominated at the revision stage, namely, equilibria in which the sender always reveals θ_H if that is indeed the state.[23]

Proposition 2. *(i) When messages are unverifiable, increasing commitment increases informativeness. (ii) When messages are verifiable, increasing commitment decreases informativeness. (iii) When $\rho = 1$, the equilibrium outcome is independent of the mode of communication: under full commitment verifiability does not matter.*

This result provides a clear set of empirical predictions that suggest experimental treatments to evaluate commitment. For both Propositions 1 and 2, we find it particularly useful that we have contrasting predictions for the different verifiability scenarios. The fact that verifiability should not matter with full commitment offers a clear prediction that we will also look to test in our experiment. The reason why this prediction comes about is that, with full commitment, under verifiability the sender is able to exactly replicate the strategy she uses when communication is unverifiable: she does so by replacing the use of message r with the no message n.

We now consider another useful comparative statics result that can be tested in our environment. This keeps fixed the degree of commitment and discusses how the equilibrium changes as we vary the persuasion threshold parameter q that measures the required posterior to induce the sender to choose a_H.

Proposition 3. *For any $\rho > 0$, for both cases of verifiable and unverifiable messages, as the persuasion threshold q increases, the strategy of the sender becomes more informative.*

As the persuasion threshold increases, the sender must reveal more information in order to induce the receiver to choose the high action. Of course, for low values

[23] In the data this type of behavior is predominant.

of ρ, in the unverifiable scenario, no information can be credibly transmitted in equilibrium for any value of q because the revision stage is too dominant in the game, and the revision stage is never informative in unverifiable scenarios. However, for ρ high enough, increasing q will strictly increase the amount of information transmitted.

We conclude our discussion of the theory with some remarks on equilibrium multiplicity. We already discussed potential alternative equilibrium outcomes under verifiable information as well as our selection criterion. However, even when there is a unique equilibrium outcome, there typically are multiple strategies that generate that outcome. This multiplicity is immaterial for our experimental results, but it is useful to outline it here. Under unverifiable information, any permutation of the messages that preserves the same information content is also an equilibrium. For instance, with $rho = 1$, we can have the sender send either θ_L or θ_H with probability one when the state is θ_H and then randomize with the required probabilities between θ_L and θ_H when the state is θ_L. For ease of exposition, we focus on equilibria with natural language when discussing our experimental analysis. This is mostly what we see in the data and it is indeed more natural. Note that, in contrast with pure cheap talk models, when ρ is sufficiently high such that commitment matters, there is no requirement that senders and receivers coordinate on an interpretation of the language: at the commitment stage the choice of information structure provides the interpretation of the language chosen by the sender. Under verifiable information, a different type of multiplicity is possible. For instance, when $\rho = 0$, all equilibria generate complete information revelation, and a message θ_H following state θ_H. However, following state θ_L, any probability distribution between messages θ_L and n is part of an equilibrium, as these all lead to the posterior belief that the state is θ_L with probability one.

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, there are now three messages that the sender can choose among, $M = \{r, b, n\}$. 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. However, the strategy method is not used 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 turned out of a different color. This design choice is intended to make the revision stage as realistic and simple as possible, and to strongly mark the difference with respect to the decision at the commitment stage. In Appendix B and C, we present the instructions and we give 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 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 variations, with different degrees of commitment, $\rho \in \{0.20, 0.80, 1\}$, generating a total of six treatments. Thus, the set of treatments forms 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 the table, 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 (16 on average per session) for a total of 384 subjects. In addition to their earnings from the experiment, subjects received a \$10 show-up fee. Average earnings, including 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 we do not include the extreme cases with $\rho = 0$, and we do so for two main reasons. First, these are the only cases for which there already exists experimental evidence, as these cases

have received attention in the literature.[24] 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 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, i.e. $\rho = 0$.

Table 3 reports the equilibrium predictions for each treatment in terms of the strategies played by senders and receivers.[25] Figure 9 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, in treatments with partial commitment, the equilibrium strategy in the revision stage is independent of ρ. More precisely, at the revision stage, in treatments with verifiable information, the sender always reveals all the information, while in treatments with unverifiable information, the sender never reveals any information. Second, 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 $U100$. In contrast, in treatment $V80$, senders are predicted to compensate by committing to reveal less information than in $V100$. Third, in both $V20$ and $U20$, the sender is unable to use commitment to fully undo her anticipated behavior at the revision stage. The predicted outcome in both these treatments is identical to the case of no commitment, $\rho = 0$. Because of this, there is 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 9). However, senders achieve this by using substantially different strategies. Note also that the equilibrium of $U100$ is the same as in $U100S$, i.e., the empty message plays no important role in the U treatment.

3.2.2 Measuring Informativeness, Revisited

For treatments with $\rho = 1$, we measure informativeness in the same way as in Section 2.2, i.e., via the *theoretical* correlation coefficients ϕ and ϕ^B. Under partial

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

[25] 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. However, there are two cases (U20 and V20) where we do not fully resolve this multiplicity, and this is harmless for our analysis.

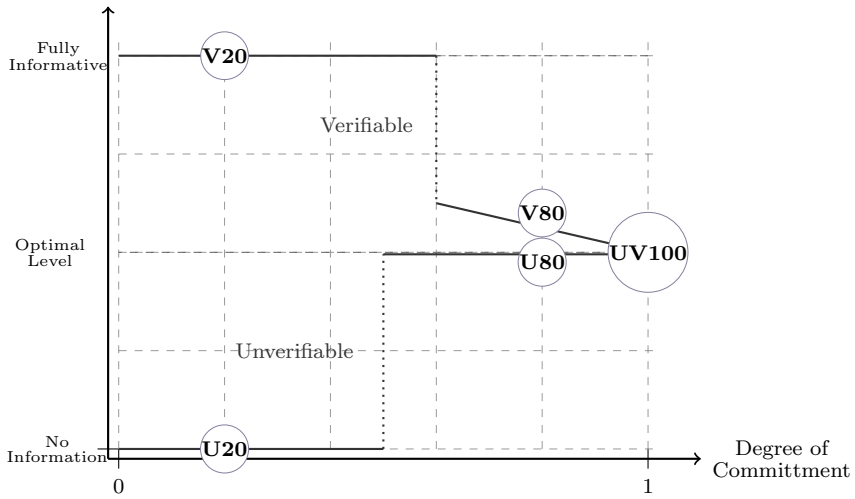


Figure 9: Predictions and Treatments.

commitment, however, the computation of the theoretical coefficient is more challenging for the following reason. For the revision stage, our design only elicits the sender's revision conditional on the *realized* state of the world θ. Therefore, we do not observe the revision that would have occurred, had the ball turned out of a different color.[26] This design choice is intended to make the revision stage, which constitutes a crucial part of our design, as realistic and straightforward as possible, and to strongly mark the difference with the commitment stage.[27] However, this choice comes with the minor drawback of missing data. We circumvent this problem by imputing the session-specific average behavior of the senders. This seems to be a natural choice and, due to random re-matching, this is what receivers should expect when facing a new sender during the experiment.[28]

4 Main Results

In this section, we present the main results of our experiment. This results generates from the framework just introduced in Section 3 and aim at uncovering how commitment and rules shape subjects behavior in the lab. More specifically, we present four sets of results.

First, we explore the simplest and most direct evidence to test whether subjects

[26]In contrast, in the commitment stage, our design collects the complete contingent plan π_C, which prescribes a message *for each* possible state. Similarly, for the guessing stage, the receiver declares a complete contingent plan, her best guess for each message (strategy method).

[27]In fact, had we used the strategy method in the revision stage, the commitment stage and the revision stage would have looked identical to the eyes of the sender.

[28]Our results are robust to different imputation methods: for example, one could impute the *subject*-specific average behavior. We note, also, that the results for treatments with $\rho = 0.8$ (where we perform such an approximation) are similar to those with $\rho = 1$ (where we do not need to use this approximation), suggesting that the results are robust to our imputation method.

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*	

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. More specifically, 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. For this role, we exploit the *across*-treatment variation and track how their responsiveness to information changes as we change the level of commitment ρ and whether it does so in ways that are consistent with the theoretical predictions.

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 again 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 or not. These asymmetric comparative statics will allow for a particularly tight test of the role of commitment in communication. More specifically, a test that can rule out possible alternative explanations for the changes we observe in the data.

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 is a particularly natural environment 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 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 represent the main departure relative to cheap talk and disclosure models. Our analysis of Section 2, despite offering important insights into subjects behavior in the presence of commitment, cannot truly uncover the effects that commitment has on subjects behavior. To test the role of commitment in communication, we need to go beyond models with full commitment and take advantage of the richer design introduced in Section 3.

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. Crucially, we do so by comparing their behavior in the commitment stage with their behavior in the revision stage. I would replace following sentences with: 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). As shown in Proposition 1, when information is unverifiable, senders should reveal substantially more information in the commitment stage than in the revision stage, and the opposite ranking should hold when information is verifiable. 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.

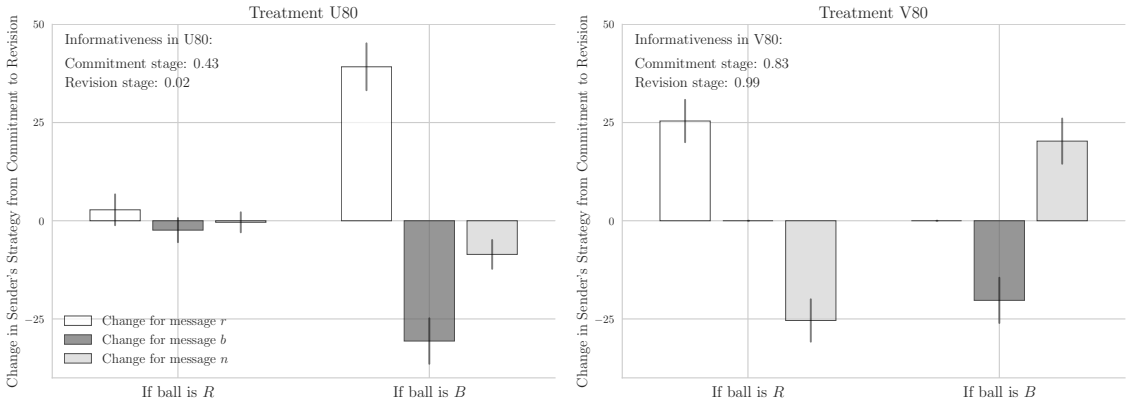


Figure 10: Sender's Strategy: Commitment vs. Revision, $\rho = 0.8$

In the left panel of Figure 10, we present the average changes that occurs in the senders' strategies, when moving from the commitment stage to the revision stages. For example, a high bar indicates a message that is used more often in the revision stage, than in the commitment stage. Overall, we find a strong qualitative change in the way senders communicate to receivers, especially so when the state is B. These changes are jointly statistically different from zero both when the state is R and when it is B (for both $p < 0.01$).[29] Altogether, these indicate that senders react to the presence of commitment. More importantly, they do so in ways that are consistent with the theoretical prediction of Table 3. The theory predicts that, when going from the first to the second stage of the game, senders should increase the total probability with which they send message r. More specifically, they should increase the probability of message r conditional on state B, while leaving it constant conditional on state R. The data of Figure 10 (left panel) are in line with this prediction. Indeed, the average total probability of message r increases to 83% (revision) from 55% (commitment). In contrast, the total probability of message b decreases to 9% (revision) from 31% (commitment). An equivalent way to confirm that the observed changes in behavior are in kine with the theoretical prediction is to look at the change in informativeness between the commitment strategy and the revision strategy. Denote these two quantities as ϕ_C^B and ϕ_R^B, respectively. In line with the prediction of Proposition 1, 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.

The evidence generated from treatment $U80$ is thus consistent with subjects taking advantage of commitment. Yet, alternative explanations, beside recognizing the rolse of commitment, are possible. For example, senders could have an intrinsic preference for being truthful in the commitment stage: not for strategic reasons, but rather because in this stage receivers can monitor their behavior. To rule out this and other alternative explanations, we can take advantage of the asymmetric prediction of our design. In treatment $V80$, in fact, senders are predicted to strategically reveal *less* information in the commitment stage, in anticipation of their behavior in the revision stage (Proposition 1). This pattern is inconsistent with the alternative explanation provided above. Consistently with the theoretical prediction, we observe that 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 $p < 0.01$). The right panel of Figure 10 illustrates the related change in senders' strategies. Most notably, we find a substantial increase, of about 25 percentage points, in the probability of sending the message r (statistically significant at the level $p < 0.01$). Recall that in treatments with verifiable information, both messages r and b are fully revealing.

[29] Although there are statistically significant changes in the same direction when the state is R, those are small in magnitude.

This suggests that senders strategically withhold message r in the commitment stage, so to increase the persuasive power of the empty message n, as predicted by the theory.

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, on average, 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 may fall short of being optimal, as we discuss in more detail later in this section. Qualitatively, however, this central prediction of our strategic communication model is corroborated by the data.[30]

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. In order to do so, we create a direct test that is specifically tailored to the problem they face. Consider the Bayesian posterior computed conditional on a message m, if we only use the information contained in the commitment strategy π_C.[31] This posterior belief, that we shall call *interim posterior*, can be interpreted as the belief that a receiver would hold if she ignored that a revision stage existed. 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. 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 11. 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 this message as a function of the induced interim posterior and the treatment.[32] On the one hand, in $U20$, the interim posterior should have little or no impact on the receiver's guess. This is 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. On the other hand, in $U100$, the interim posterior should have a

[30] In Appendix **??**, we perform the same analysis on treatments $U20$ and $V20$. coming to similar conclusions.

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

[32] In Figure 11 the solid lines are the polynomial fit of the induced interim posterior and the observed guess.

substantial effect on the receiver's guess. In particular, the receiver should guess *red* for high enough posteriors. In fact, since 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 11 is mostly flat in $U20$, whereas it is strictly increasing in $U100$.

A similar, if not stronger, evidence is found when comparing $V20$ and $V100$. For this test, however, the *strategic message* to consider is n, the empty message. By the nature of verifiable information, indeed, messages r and b cannot be used as they induce trivial interim beliefs of either 1 or 0, respectively. The message that potentially entails strategic considerations is message n, and this is what we focus on. As can be seen in the right panel of Figure 11, the estimated receivers' response to an increase in the interim posterior is weak in treatment $V20$, whereas it is strong and positive for $V100$.[33] Overall, the joint evidence coming from treatments with verifiable and with unverifiable information suggests that 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.

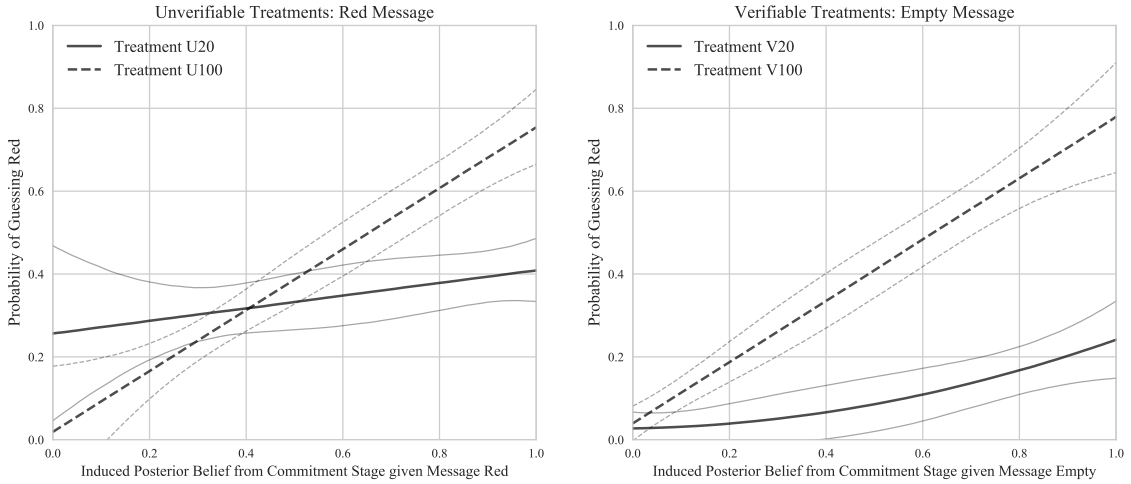


Figure 11: Receiver's Response to Persuasive Messages: $\rho = 0.2$ vs. $\rho = 1$

4.2 Cross-Treatment Comparisons

The previous section establishes that both senders and receivers react to the power of commitment. This section explores more explicitly the extent to which such re-

[33]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 $\rho = 1$. This is true 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 $\frac{1}{2}$, there is a statistically significant difference in both cases ($p < 0.01$ in both cases) and, maybe more importantly, the magnitude of the change is much more important: 56 versus 14 percentage points in the verifiable case, and 40 versus 6 percentage points in the unverifiable case.

action goes indeed in the directions suggested by the theory. One key prediction of the theory, stated in Proposition 2, concerns how equilibrium informativeness should change with commitment under verifiable and unverifiable information. Figure 12 shows the distributions of sender-specific informativeness ϕ^B in our main treatments. There are two main takeaways from this figure. On the one hand, there is 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 $U20$. This suggests that under unverifiable information, the amount of information transmitted by the senders increases as commitment increases, as predicted by the theory. On the other hand, there is a first-order stochastic *decrease* in the distribution of informativeness in $V100$ relative to $V20$ (right panel), and less so in $V80$ relative to $V20$. This suggests that, under verifiable information, the amount of information transmitted by the senders decreases as commitment increases, as predicted by the theory.[34]

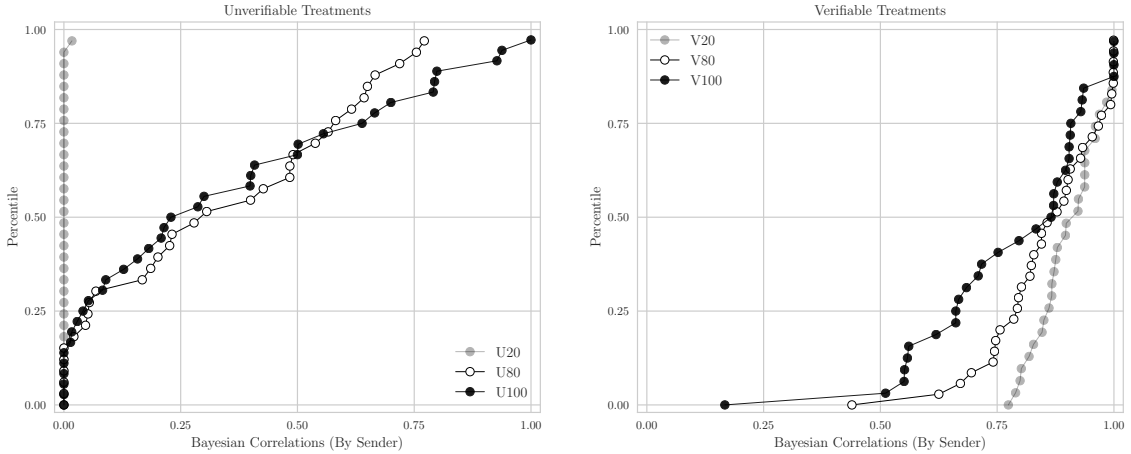


Figure 12: CDF of Subject Average Bayes Correlation ($\bar{\phi}_i^B$) by Treatment

While these changes are qualitatively in line with theory, there are important quantitative deviations from the point-predictions of the theory that are worth noting. To illustrate this, instead of looking at the distributions of informativeness for different treatments and how they compare with each other, we compare 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. Considering the differences between ϕ^B and ϕ allows us to partly disentangle whether it is senders or receivers who are mainly responsible for possible deviations from the equilibrium. The symbols between the numbers indicate their

[34] 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 in passing 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 D30 in the appendix.

"statistical relations" at the 10% level, i.e. $\approx$ means that the p-value of the equality of the two is larger than 0.1.

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.

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, the observed changes are more muted than what is predicted by the theory. In the case of verifiable information, for example, the change from $V20$ to $V100$ is predicted to reduce the correlation from 1 to 0.5; but the difference in the data is 0.13, or only the 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 that only part of the "missing effect" can be imputed to mistakes on the receivers' side. Recall that when we replace our actual receivers with a hypothetical Bayesian receiver, we are effectively shutting 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 a 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 treatments. However, it is also clear, especially for the unverifiable treatments, that many sender's have correlations that are positive, but low enough to suggest that some of their strategies must not move the posteriors enough.

To understand this better we now turn to the analysis of what posteriors senders induce with their communication strategies. In particular, Figure 13 displays the kernel density estimates of the Bayesian posteriors *conditional on the state*.[35] The vertical dashed lines indicate the theoretical predictions; the other lines present

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

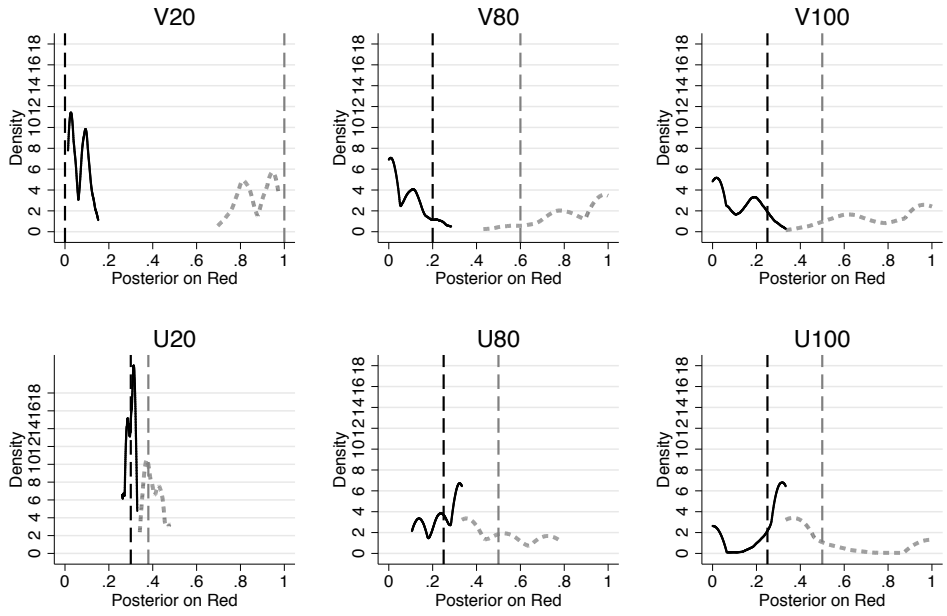


Figure 13: Posterior on R as a Function of the State

the data under the different treatments.[36] 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 there is 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, while moving from V20 to V100, the posteriors move closer, as predicted by theory. However, there are some important discrepancies: While in the case of V20 the posteriors are inside the 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 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 makes it clear that the data moves in the right direction for both verifiable and unverifiable treatments, but that mean difference is much closer to the theoretical predictions in the case of

[36] In the bottom left panel, the two vertical lines are not both at one third 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 (1.00)		0.78 (0.40)		0.69 (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 (0.00)		0.24 (0.25)		0.30 (0.25)	
	B	***R***	***B***	***R***	***B***	***R***
Mean:	0.30	0.41	0.25	0.49	0.23	0.53

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. Remember that in theory, the two treatments with $\rho = 1$ should yield equivalent behavior.

At this point, we have the following takeaways: The behavior of our subjects is not, on average, in line with the point predictions of the theory. However, our data qualitatively matches 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

Although this is not the focus of the preceding section, the results suggests that the impact of rules when commitment is low is in line with the theoretical prediction. Indeed, this is the case, as can be seen in Figure 14. At $\rho = 0.2$, senders convey much more information when messages are verifiable. Correlations with Bayesian receivers are close to the predicted values. The impact on receivers is not as large, but is nonetheless substantial. This can be seen in the right panel, which indicates the probability of guessing *red* after receiving a r message. Under verifiable information the probability is close to one (0.99), while it is substantially less at 0.35 under unverifiable messages. This also makes it clear that some receivers follow the sender's

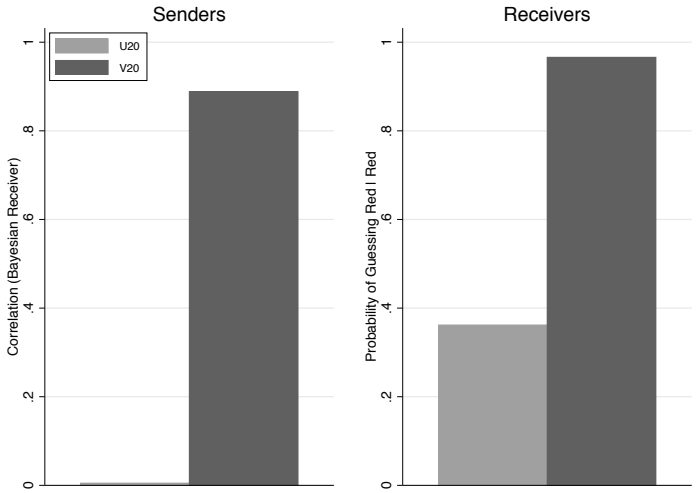


Figure 14: Impact of Rules Without Commitment

recommendation despite the posterior being below 0.5.[37]

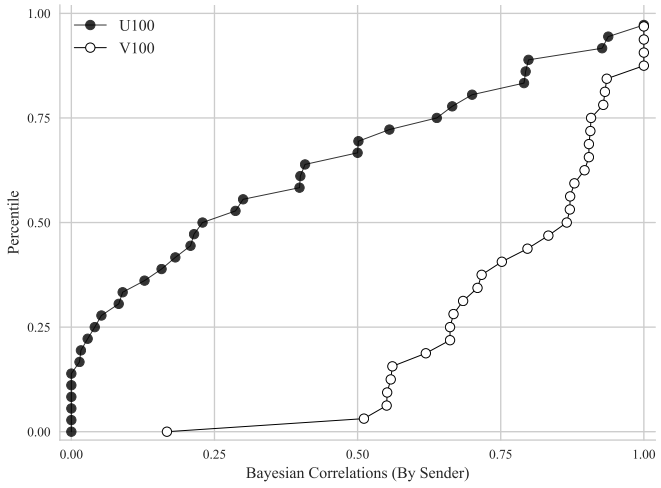


Figure 15: CDF of Subject Average Bayesian Correlation ($\bar{\phi}_i^B$): V100 and U100

At the other extreme, there remains an unpredicted impact of rules when commitment power is high. Figure 15 shows the CDFs of sender average Bayesian correlation for treatments U100 and V100. It makes it clear that more information is transmitted when rules are verifiable.

As we saw in the previous section, this difference seems (at least in part) driven by the fact that in the V100 treatment, the difference in posteriors conditional on

[37]This is the counterpart, in our setting, of the common finding from cheap-talk experiments that there are receivers that follow the sender's recommendation when they should not. Figures D24, D25, D26, and D28 in Appendix D provide more details about receivers behavior for the general treatments.

Table 6: Theoretical Predictions and Data: V100 and U100

U100

Theory:

		Messages		
		r	b	n
States	R	100%	0	0
	B	50%	50%	0

Data:

		Messages		
		r	b	n
States	R	74%	13%	13%
	B	48%	38%	14%

V100

		Messages		
		r	b	n
States	R	0	0	100%
	B	0	50%	50%

		Messages		
		r	b	n
States	R	52%	0	48%
	B	0	55%	45%

the state is too large; much larger than predicted. What causes the difference in posteriors documented above to be so far from the theoretical prediction in the V100 treatment? Table 6, which gives the theoretical prediction and average behavior in terms of strategies hints at part of the explanation. Senders in the V100 treatment ought to always hide "good news" in order to make it possible to sometimes hide the bad news. Instead, in the aggregate, subjects commit to approximately 50-50 randomization with the empty message following **both** states. This undermines the logic of using the *no* message to substitute the *red* one in constructing the optimal Bayesian persuasion strategy. In contrast, in U100, senders have to strategically lie about "bad news," and indeed the average behavior is relatively close to that. Hence, the aggregate data suggests that, on balance, subjects are less successful in implementing a persuasion strategy in the verifiable treatment.

Disaggregating sender's behavior in the V100 treatment reveals considerable heterogeneity however. Table 16 reports the clustered strategies of the senders in our sample.This shows that the most common type is in fact using a strategy that comes close to the equilibrium prediction: it mostly uses the n message when the state is R and mixes between b and n when the state is B. Notice, however, that the way in which that group deviates from equilibrium moves it toward more information revelation. The second most popular type of strategy involves revealing the "good news" and trying to hide the "bad news." Hence, this should result in complete information transmission, but it suggests that the senders are trying to be strategic. The third important cluster can be best described as "revealing:" the sender mostly indicate the state through their messages.

To summarize, rules have an impact that the theory does not predict when $\rho = 1$, but a substantial fraction of senders are actually displaying behavior close to what

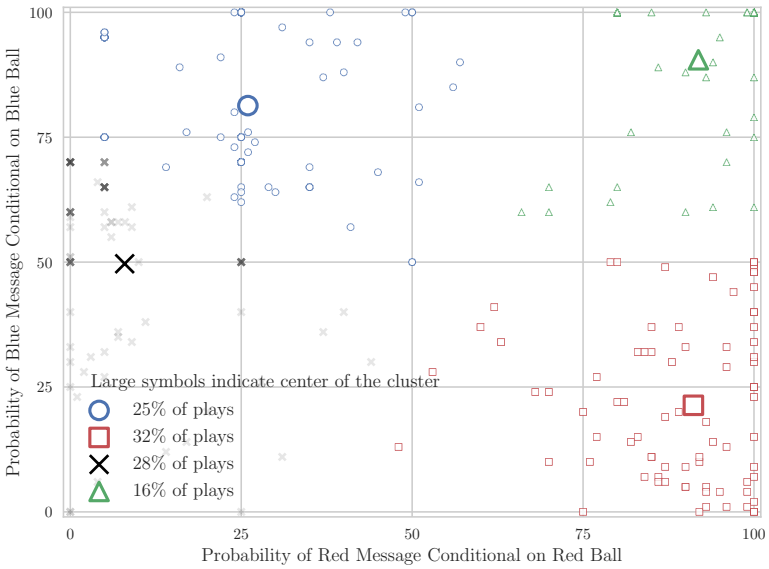


Figure 16: Sender's Strategy Grouped in Clusters for Treatment V100

the equilibrium suggests.

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 should lead to more informative communication from senders. Based on this idea, we designed one more treatment, where it is more valuable for the receiver to match the state when the state is B. This 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 with the treatments above, the receiver wins different amounts if he correctly guesses the color of the ball: 2 if 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 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 to 20 subjects (72 in total). Those subjects made between \$10.48 and \$26.56 (average \$18.02).

Figure 17 shows that 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,

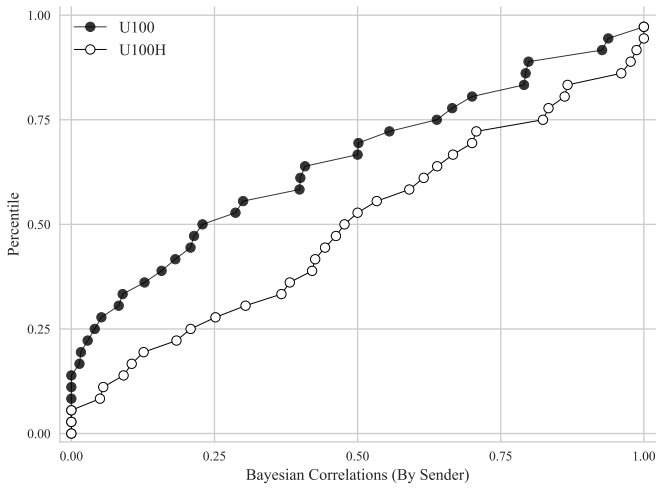


Figure 17: CDF of Subject Average Bayesian Correlation ($\bar{\phi}_i^B$): U100 and U100H

is not statistically significant ($p > 0.1$).[38] We note though, that sender behavior in the U100H treatment evolves a lot, as can be seen in Figure D21, 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-interraction variable is significant and positive for the U100H treatment ($p < 0.01$).[39] Hence, a wedge is indeed building over time, with senders ultimately conveying more information in the U100H treatment than the U100 treatment.

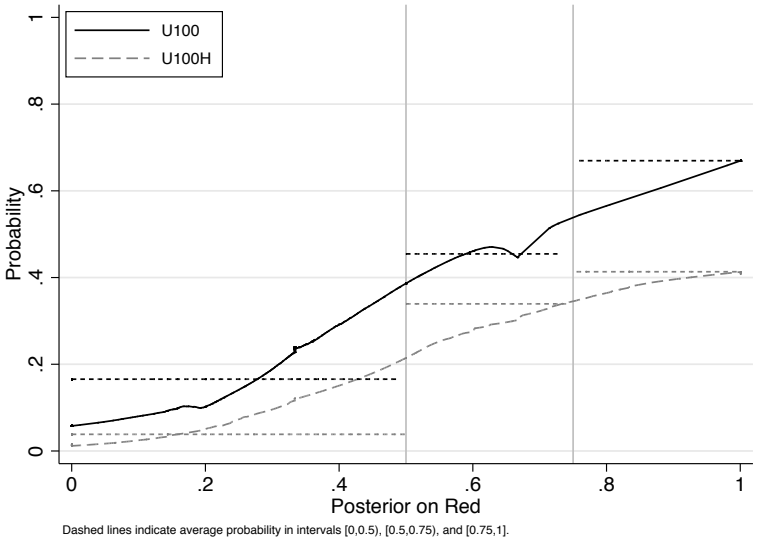


Figure 18: Probability of Guessing *red* by Posterior for treatments U100 and U100H

[38] t-test clustering at session level.

[39] In order to account for time, this test does not average at the subject level.

The theory suggests that senders need to send more information to convince receivers to listen to them. Figure 18 shows that 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 ($p > 0.1$).[40]

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 where 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 babling equilibrium of cheap-talk games. However, when 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 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, one more way to examine whether subjects react to commitment as predicted. Finally our experiment provides one of the first experimental investigation of Kamenica & Gentzkow (2011).[41]

Our findings suggest that the central force at play in Kamenica & Gentzkow (2011) is one that many subjects recognize. Indeed most aspects of aggregate behavior are in line with the qualitative predictions of Kamenica & 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 & Gentzkow (2011) model seems to be ones that subjects react to despite not being perfectly rational and optimizing

[40] Appendix D provides additional information on the behavior of receivers in Figure D29.

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

agents. In that sense it offers a useful starting point to model communication in the presence of commitment. Our extension to partial commitment offers both 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.

References

Abu-Mostafa, Y. S., Magdon-Ismail, M. & Lin, H.-T. (2012), *Learning from data*, Vol. 4, AMLBook New York, NY, USA:.

Alonso, R. & Camara, O. (2016), 'Persuading voters', *The American Economic Review* **106**(11), 3590–3605(16).

Au, P. H. & Li, K. K. (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.

Battaglini, M. (2002), 'Multiple referrals and multidimensional cheap talk', *Econometrica* **70(4)**, 1379–1401.

Benndorf, V., Kübler, D. & Normann, H.-T. (2015), 'Privacy concerns, voluntary disclosure of information, and unraveling: An experiment', *European Economic Review* **75**, 43–59.

Blume, A., De Jong, D., Kim, Y. & Sprinkle, G. (1998), 'Experimental Evidence on the Evolution of Meaning of Messages in Sender-Receiver Games', *The American Economic Review* **88**(5), 1323–1340.

Blume, A., Lai, E. K. & Lim, W. (2017), 'Strategic information transmission: A survey of experiments and theoretical foundations', *Working Paper* .

Bock, O., Baetge, I. & Nicklisch, A. (2014), 'hroot: Hamburg registration and organization online tool', *European Economic Review* **71**, 117–120.

Cai, H. & Wang, J. T. Y. (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**(2), 163–183.

Charness, G. & Levin, D. (2005), 'When optimal choices feel wrong: A laboratory study of bayesian updating, complexity, and affect', *American Economic Review* **95**(4), 1300–1309.

Crawford, V. P. & Sobel, J. (1982), 'Strategic Information Transmission', *Econometrica* **50**, 1431–1451.

Dickhaut, J., Ledyard, M., Mukherji, A. & Sapra, H. (2003), 'Information management and valuation: an experimental investigation', *Games and Economic Behavior* **44**(1), 26–53.

Dickhaut, J., McCabe, K. & Mukherji, A. (1995), 'An Experimental Study of Strategic Information Transmission', *Economic Theory* **6**(3), 389–403.

Dranove, D. & Jin, G. (2010), 'Quality Disclosure and Certification: Theory and Practice', *Journal of Economic Literature* **48**(4), 935–963.

Dye, R. (1985), 'Disclosure of nonproprietary information', *Journal of Accounting Research* **23(1)**, 123–145.

Forsythe, R., Isaac, R. M. & Palfrey, T. R. (1989), 'Theories and Tests of "Blind Bidding" in Sealed-bid Auctions', *The RAND Journal of Economics* **20**(2), 214–238.

Forsythe, R., Lundholm, R. & Rietz, T. (1999), 'Cheap Talk, Fraud, and Adverse Selection in Financial Markets: Some Experimental Evidence', *The Review of Financial Studies* **12**(3), 481–518.

Fréchette, G. R. (2012), 'Session-effects in the laboratory', *Experimental Economics* **15**(3), 485–498.

Galor, E. (1985), 'Information sharing in oligopoly', *Econometrica* **53**, 329–343.

Gentzkow, M. & Kamenica, E. (2014), 'Costly persuasion', *American Economic Review* **104**(5), 457–462.

Gilligan, T. W. & Krehbiel, K. (1987), 'Decisionmaking and Standing Committees: An Informa-

tional Rationale for Restrictive Amendment Procedures', *Journal of Law, Economics* **3**(2), 287–335.

Gilligan, T. W. & Krehbiel, K. (1989), 'Information and Legislative Rules with a Heterogeneous Committee', *American Journal of Political Science* **33**(2), 459–490.

Gilligan, T. W. & Krehbiel, K. (2016), 'Is no news (perceived as) bad news? An experimental investigation of information disclosure', *NBER Working Paper* .

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., Koessler, F. & Perez-Richet, E. (2014), 'Certifiable pre-play communication: Full disclosure', *Econometrica* **82(3)**, 1093–1131.

Hagenback, J. & Perez-Richet, E. (2018), 'Communication with evidence in the lab', *Working Paper* .

Hastie, T., Tibshirani, R. & Friedman, J. (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, & strategic behavior*, Pearson Addison Wesley Boston, MA.

Jin, G. & Leslie, P. (2003), 'The effect of information on product quality: Evidence from restaurant hygiene grade cards', *The Quarterly Journal of Economics* .

Jin, G., Luca, M. & Martin, D. (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. **URL:** *http://www.jstor.org/view/0361915x/di010140/01p0004l/0*

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

King, R. & Wallin, D. (1991), 'Market-induced information disclosures: An experimental markets investigation ', *Contemporary Accounting Research* **8**(1), 170197.

MacQueen, J. (1967), Some methods for classification and analysis of multivariate observations, *in* 'Proceedings of the fifth Berkeley symposium on mathematical statistics and probability', Vol. 1, Oakland, CA, USA, pp. 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**(2), 380–391.

Min, D. (2017), 'Bayesian persuasion under partial commitment', *Working Paper* .

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

Nguyen, Q. (2017), 'Bayesian persuasion: Evidence from the laboratory', *Working Paper* .

Okuno-Fujiwara, M., Postlewaite, A. & Suzumura, K. (1990), 'Strategic Information Revelation', *The Review of Economic Studies* **57**, 25–47.

Sánchez-Pagés, S. & Vorsatz, M. (2007), 'An experimental study of truth-telling in a sender-receiver game', *Games and Economic Behavior* **61**(1), 86–112.

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

Wang, J., Spezio, M. & Camerer, C. (2010), 'Pinocchio's pupil: Using eyetracking and pupil dilation to understand truth telling and deception in sender-receiver games', *The American Economic Review* **100**(3), 984–1007.

Wilson, A. & Vespa, E. (2017), 'Information transmission under the shadow of the future: An experiment', *Working Paper* .

Online Appendix for

Rules and Commitment in Communication

Guillaume Fréchette
NYU

Alessandro Lizzeri
NYU

Jacopo Perego
Columbia

Contents

A Proofs (Incomplete)

A.1 Proof of Proposition 1

Part (i) We first consider the unverifiable scenario.

(1) We first show that, if the receiver chooses the action a_H with positive probability, then information transmission can only take place at the commitment stage. Because $\mu < q$, for the receiver to choose a_H with positive probability, there must be a message m_1 that induces a higher posterior for the receiver than some other message m_2, which in turn must induce a posterior below μ. Message m_1 induces the receiver to choose a_H with positive probability while m_2 induces the receiver to choose a_L with probability one. This implies then the sender always chooses m_1 in the revision stage.

(2) We now establish that there is a threshold $0 < \widehat{\rho} < 1$ such that, for $\rho > \widehat{\rho}$ the receiver chooses a_H with positive probability in equilibrium. First note that, because $\mu < q$, the receiver must receive some information in order to choose a_H. Thus, the sender would like to use commitment to transmit information. If some information is transmitted in the revision stage, then it is clear that positive information is transmitted. Assume then that no information is transmitted in the revision stage. Consider a strategy by the sender that chooses full revelation in the commitment stage, for instance, by choosing an information structure of θ_H with probability one if the state is θ_H and θ_L with probability one if the state is θ_L. If ρ is sufficiently high, then the posterior induced by this strategy following a message of θ_H is larger than μ, so the receiver is happy to choose a_H following such a strategy and message. Therefore, for sufficiently high ρ the sender will choose to transmit a positive amount of information and induce the receiver to choose a_H with positive probability.

Consider now the verifiable scenario.

(3) First note that any undominated strategy in the revision stage involves disclosing the high state. All these strategies in the revision stage imply full information revelation at this stage.

(4) Now, recall that verifiability implies that, if the state is θ_H the sender can either send θ_H or n but cannot send θ_L. Analogously, if the state is θ_L, the sender cannot send θ_H. Thus, in order to attempt to maximally persuade the receiver, the commitment stage must involve using the n message when the state is θ_H and randomizing between θ_L and n when the state is θ_L. However, if observing n induces the receiver to choose a_H, then, at the revision stage, the sender optimally chooses n with probability 1 if the state is θ_L. If ρ is low, then this means that there is no way that a choice of n in the commitment stage leads the receiver to choose a_H. Thus, this leads the receiver to choose a_L following both message n and message θ_L. As a result, this strategy at the commitment stage leads to lower payoffs for

the sender than the outcome of full information revelation that involves sending θ_H with probability one in the commitment stage. This means that, for low values of ρ, the equilibrium involves full information revelation. Now, consider a ρ close to 1. Assume that the sender at the commitment stage chooses n with probability one if the state is θ_H, and θ_L with probability one when the state is θ_L, then, even if the sender chooses n in the revision stage when the state is θ_L, this is not enough to overwhelm the information contained in n from the commitment stage. Thus, in this scenario, following message n, the receiver chooses a_H. In fact, because the receiver now chooses a_H with positive probability even when the state is θ_L (following the choice of n in the revision stage), the overall payoff of this specific commitment strategy for the sender is higher than the one obtained under full disclosure, proving that, for high enough ρ, it is beneficial for the sender to choose a strategy that leads to lower information transmission. In fact, for high enough ρ, at the commitment stage the sender can choose to randomize between n and θ_L when the state is θ_L to induce and even higher probability of choice of a_H.

Part (ii) This is an immediate consequences of the comparison between steps 1 and 2 for the unverifiable case, and steps 3 and 4 for the verifiable case.

□

A.2 Proof of Proposition 2

We first argue that, when $\rho = 1$, the equilibrium outcome is independent of verifiability. In both scenarios, the sender wishes to maximize the probability that the receiver chooses a_H, subject to the Bayes' constraint that the average posterior must equal the prior. The optimal strategy therefore must involve the selection of a persuasive message that leads the receiver to choose a_H while being indifferent between a_H and a_L. Thus, the posterior of the receiver upon receiving the persuasive message must be equal to the prior, whereas the posterior following any other message must be equal to zero. The following strategy implements this optimum in the unverifiable scenario: if the state is θ_H send message θ_H with probability one; if state is θ_L randomize between θ_H and θ_L in such a way that the posterior following a message of θ_H is equal to q. In the verifiable case, this strategy is not feasible, but the following alternative strategy by the sender implements the same beliefs and actions for the receiver: if state is θ_H send message n with probability one; if state is θ_L randomize between n and θ_L in such a way that the posterior following a message of θ_H is equal to q.

When ρ decreases, the revision stage imposes constraints on what the sender can achieve with the commitment strategy. As shown in the proof of Proposition 1, the strategy in the revision stage is independent of ρ for both verifiable and unverifiable scenarios, with no information transmitted in the unverifiable scenario

and full information transmission in the verifiable scenario. Thus, if the strategy at the commitment stage is left unchanged as ρ decreases, the revision stage leads to overall less information communicated for lower ρ with unverifiable information and more information with verifiable information. The sender will therefore modify the commitment strategy to attempt to minimize the effect of the revision stage. As ρ decreases, the sender will reduce the amount of information transmitted at the commitment stage in the verifiable case and increase it in the unverifiable case. However, the smaller is ρ, the less the sender is able to do this. Since this constraint monotonically becomes tighter as ρ decreases, the amount of information that is communicated overall between the commitment and revision stages also changes monotonically.□

A.3 Proof of Proposition 3

Note first that changing q has no effect on the equilibrium strategies in the revision stage: as shown in Proposition 1, there is full revelation in the verifiable case and no revelation in the unverifiable case. Now recall that q equals the minimal posterior for the receiver to choose a_H. As q increases, the sender has less room for randomizing between messages in the low state while still maintaining incentives for the receiver. For $\rho = 1$, the result is immediate since the equilibrium posterior conditional on the persuasive message must equal q. Thus, as q increases, the sender reduces the probability of sending the persuasive message when the state is θ_L, and therefore reveals more information. For $\rho < 1$, the logic is similar: in both verifiable and unverifiable cases the sender uses the commitment stage to undo the outcome in the revision stage as much as possible to maintain the incentives of the receiver. As q increases, this involves revealing more information.

B Design

Figures B19 and B20 show the relevant screenshots from our experiment. The top panel of Figure B19 shows the sender's decision screens. Figure B20 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 B20 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 B19: Screens 1 and 2, Treatment U80

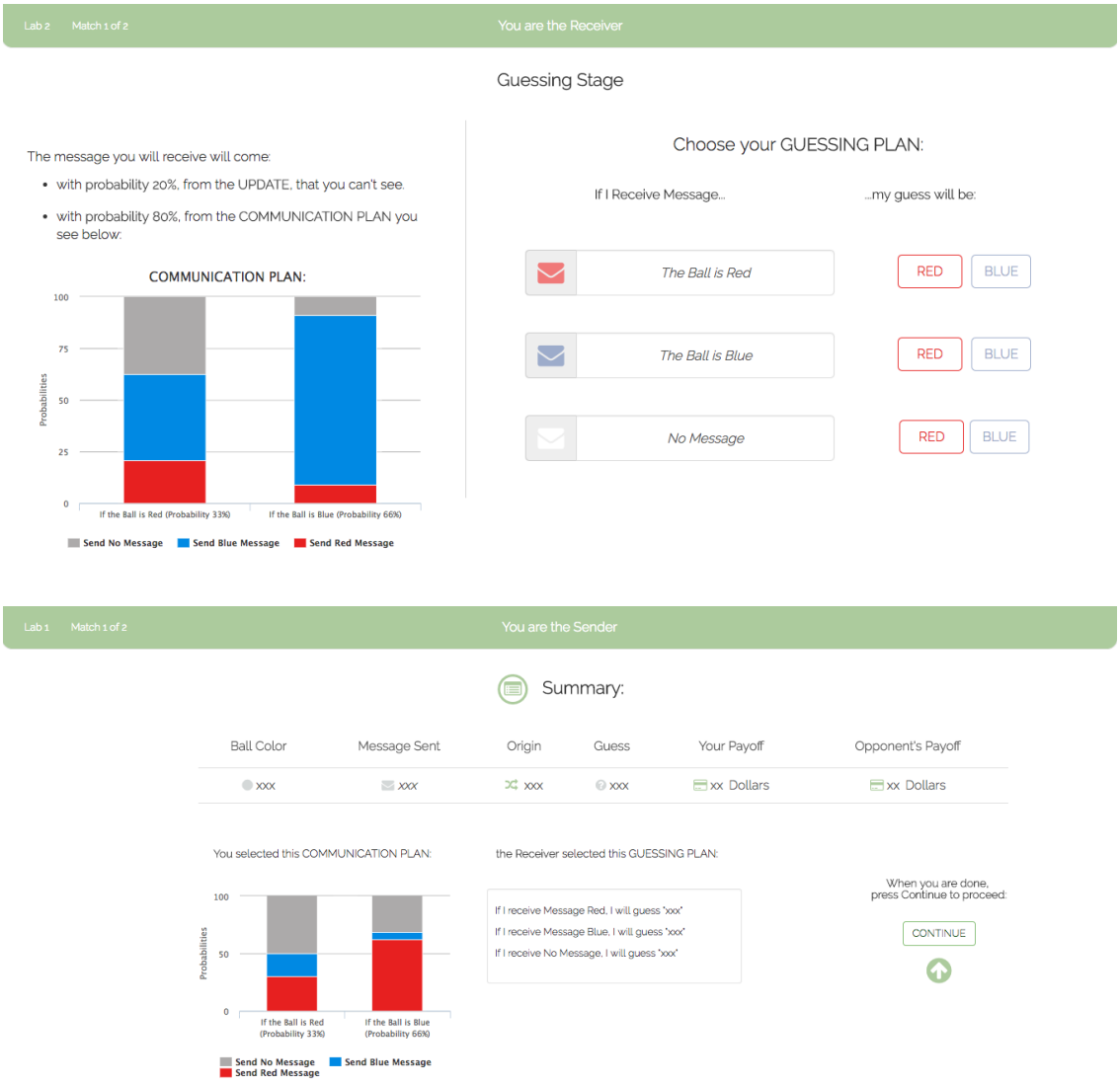


Figure B20: Screens 3 and 4, Treatment U80

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 C7.

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 C8.

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.

	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 C7: Payoffs

If Ball is Red:	If Ball is Blue:
– Message: "*The Ball is Red.*"	– Message: "*The Ball is Blue.*"
– No Message.	– No Message.

Table C8: 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

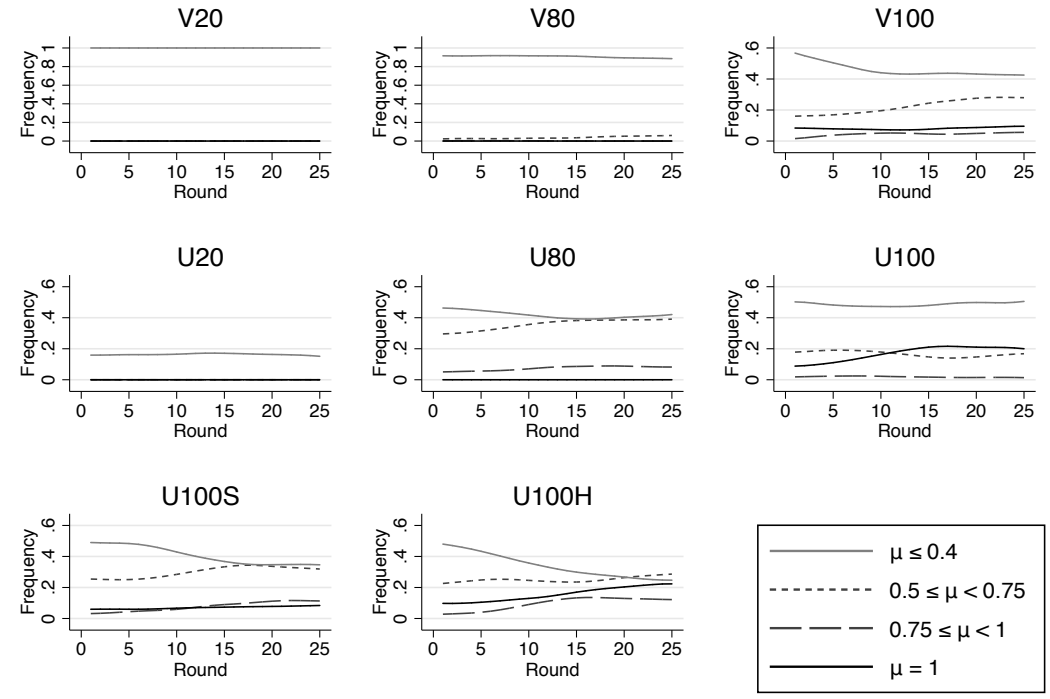


Figure D21: Frequency of Persuasive Messages Grouped by Posterior (μ)

The next two figures illustrate changes in behavior over the course of the experiment. Figure D21 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.

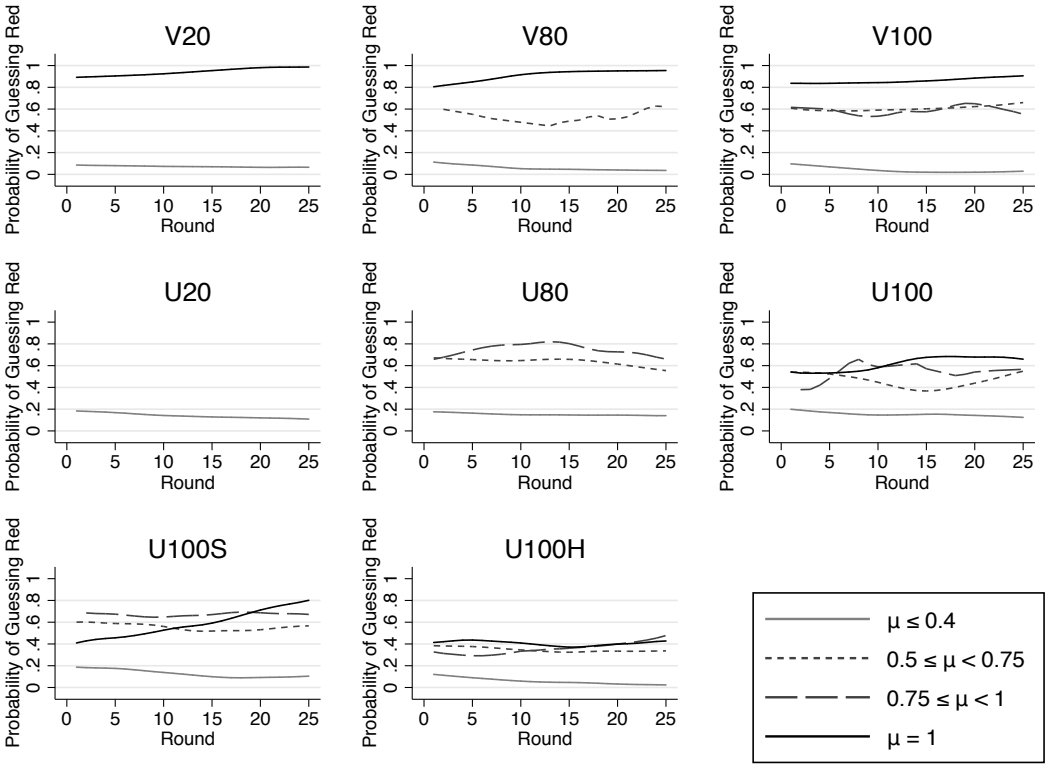


Figure D22: Frequency of Guessing *red* Grouped by Posterior (μ)

On the responder side, Figure D22 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.

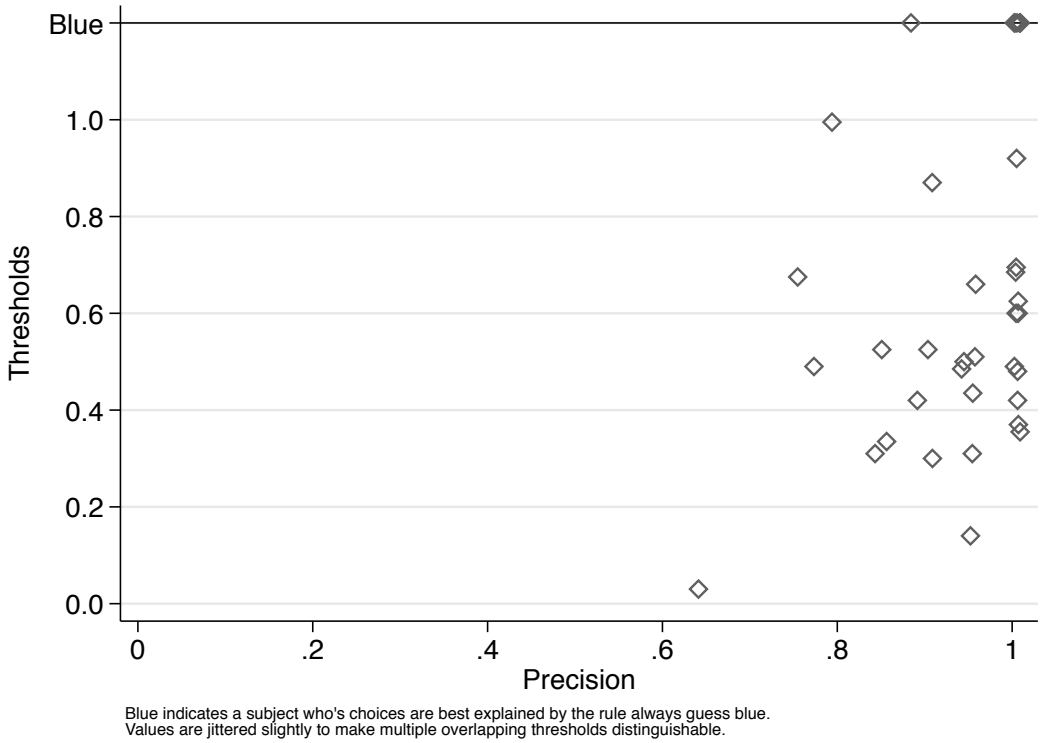


Figure D23: Estimated Threshold and Precision for Treatment U100S

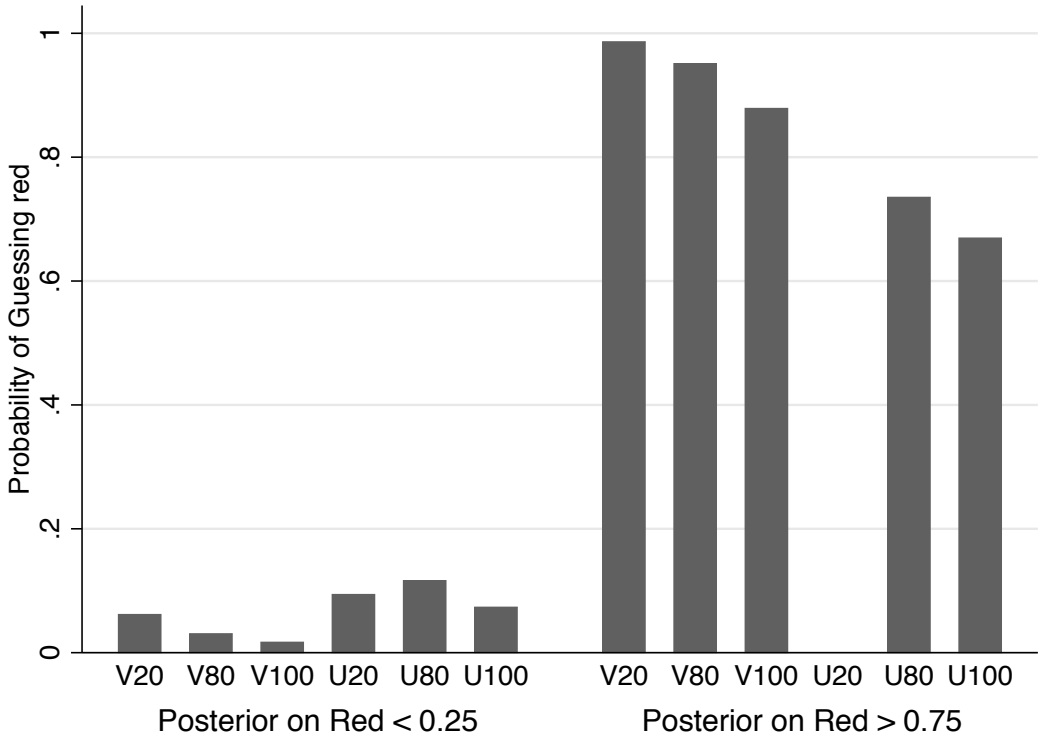
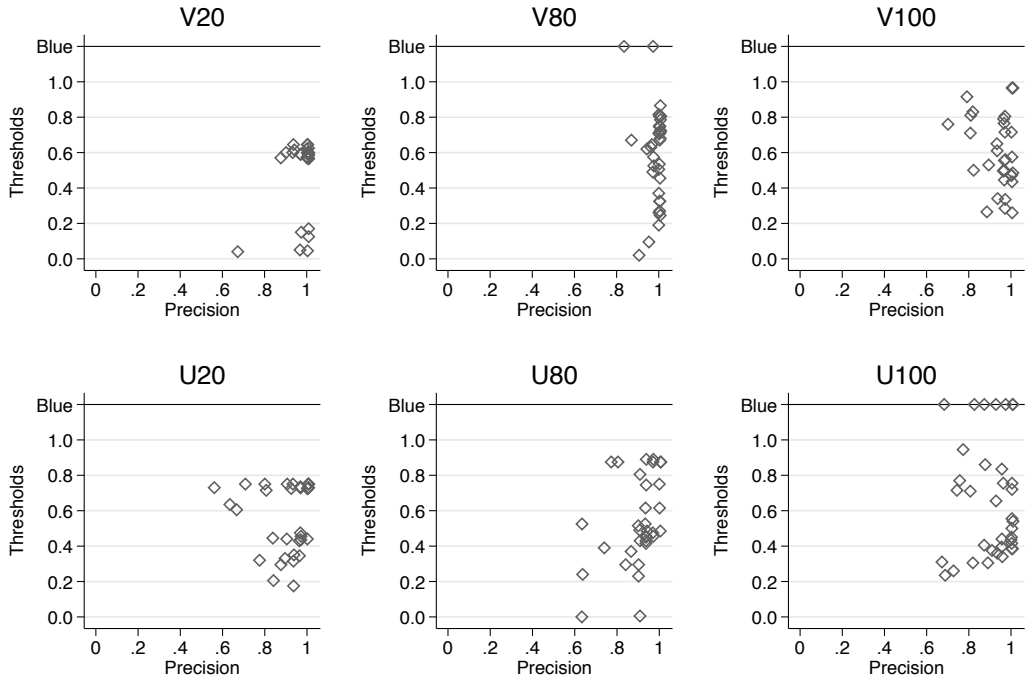


Figure D24: Probability of Guessing Red as a Function of Posterior

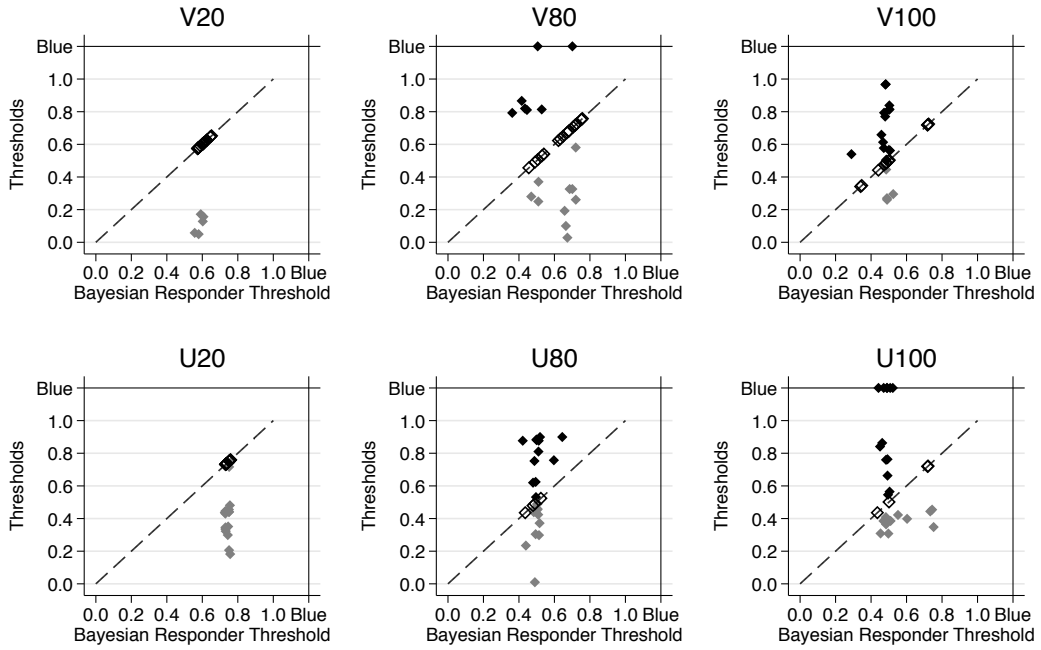
Voting patterns in all treatments are similar in that they are increasing in the posterior on red. They do display some revealing differences however. As can be seen in Figure D24, treatments with verifiable messages lead to more "certainty" in voting. When the posterior on red is low, the probability 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 D25: Estimated Threshold and Precision

Figure D25 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.



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 D26: Estimated Threshold: Actual Receivers Against Bayesian

Figure D26 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 2.3, 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).

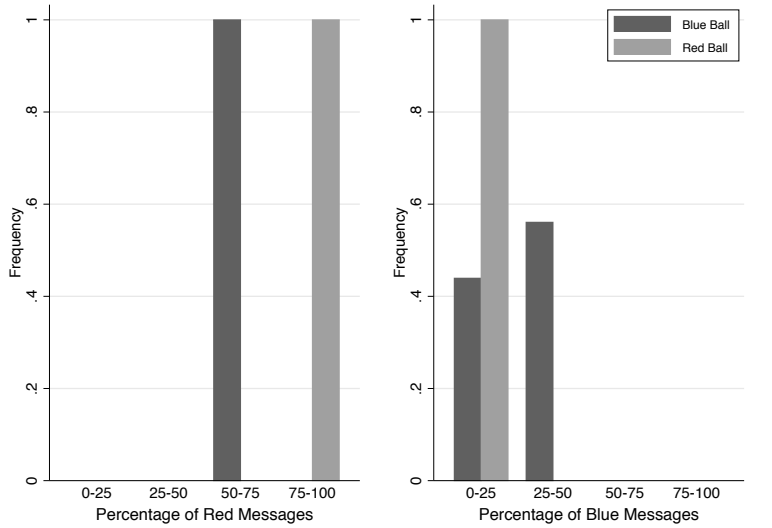


Figure D27: U20 Revision Stage

The parallel to point 2 can be evaluated in Figure D27. 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 D28 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.

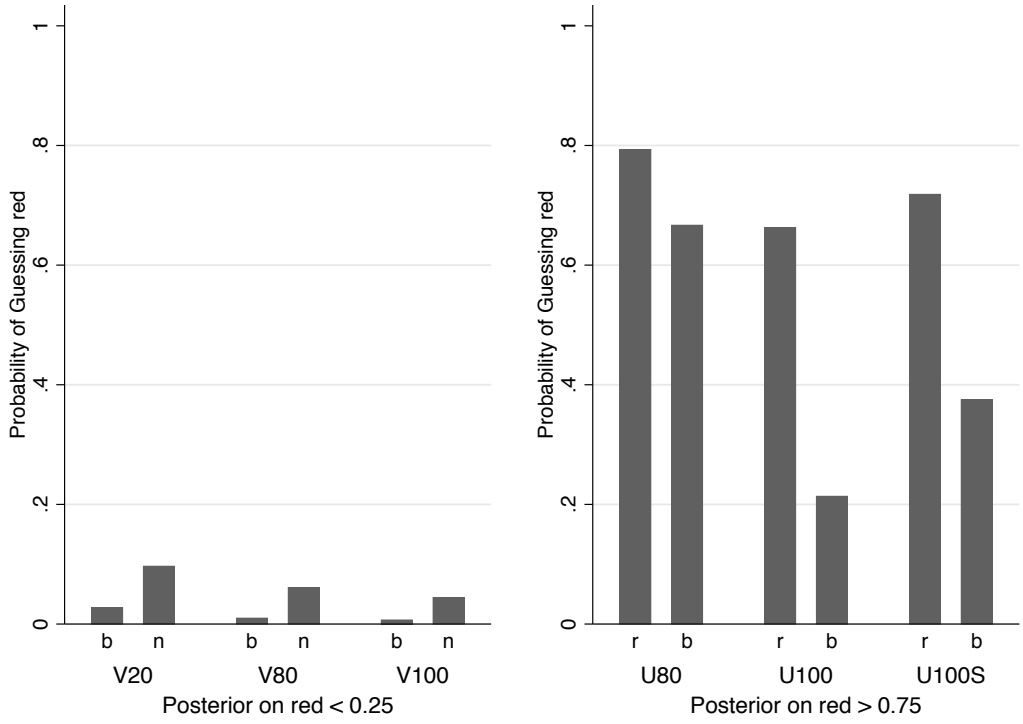


Figure D28: Probability of Guessing *red* as a Function of the Message

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 D28, 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.

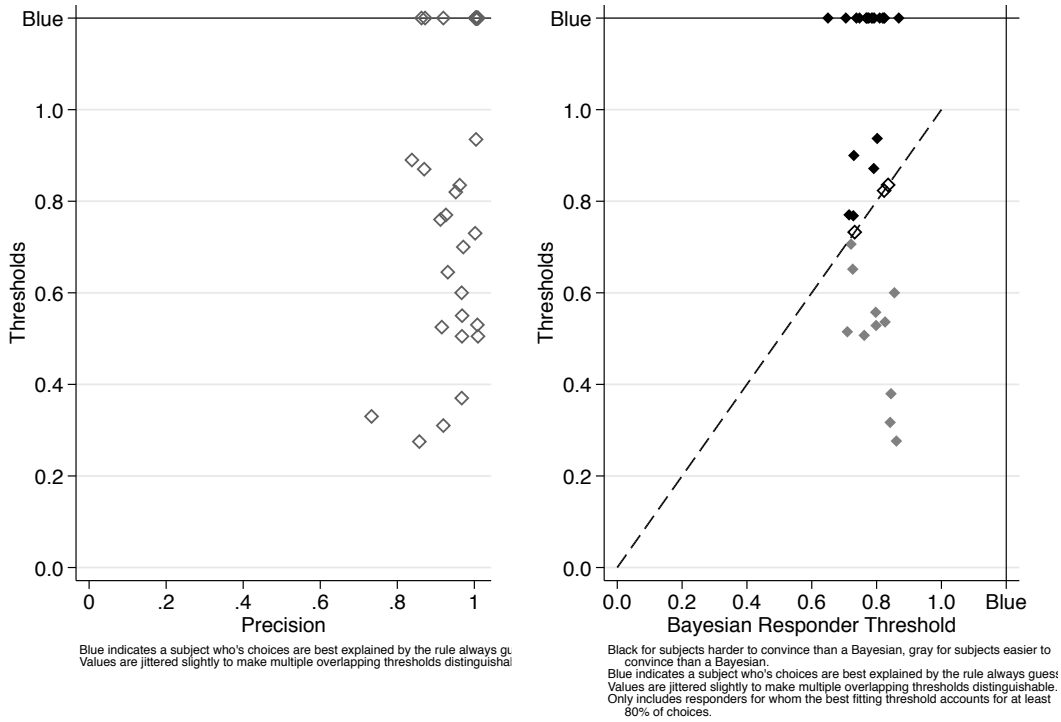


Figure D29: Estimated Threshold and Precision: U100H

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 D29.

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 D29 shows there is heterogeneity in terms of how thresholds compare to what Bayesian receivers would do.

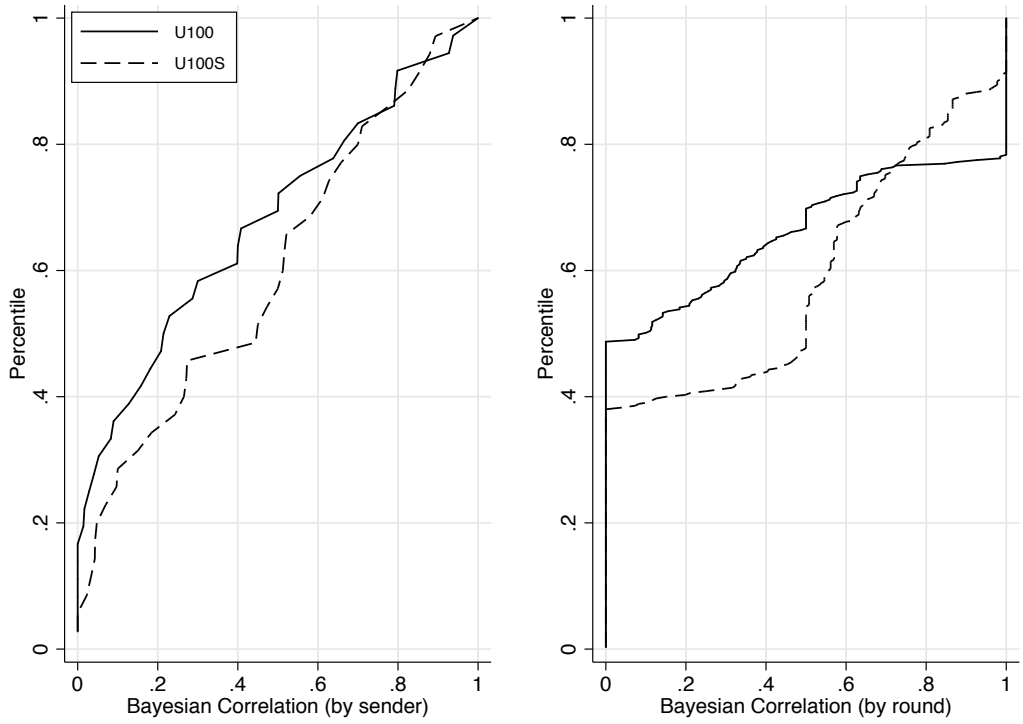


Figure D30: CDF of Bayes Correlation (ϕ^B): U100 and U100S

Figure D30 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 D30 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.

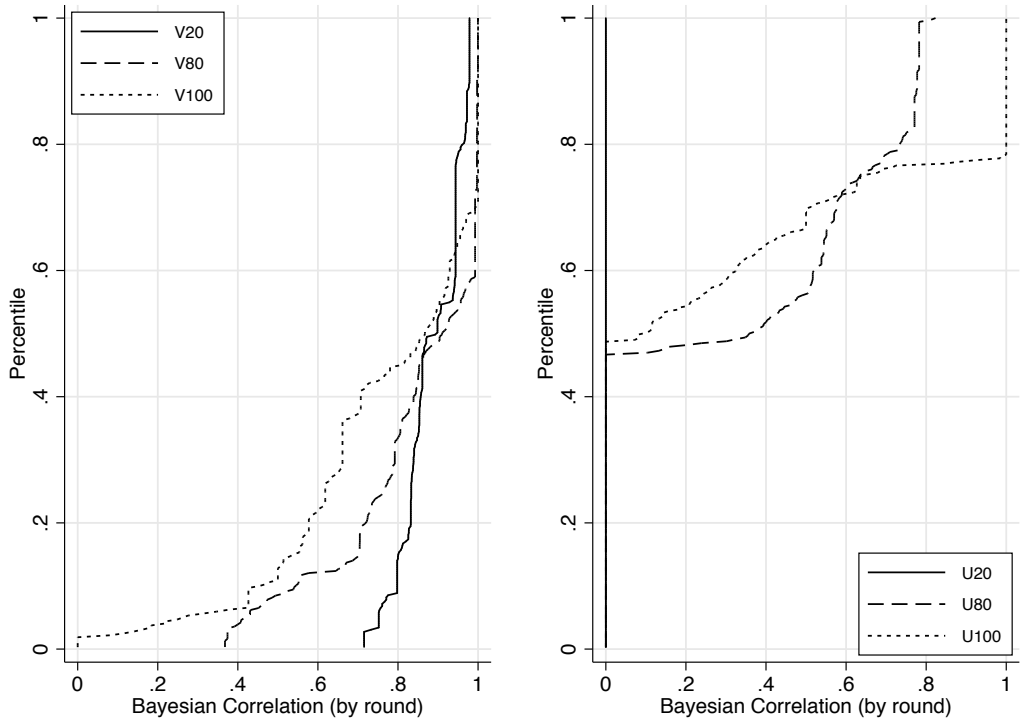


Figure D31: CDF of Bayes Correlation (ϕ^B)

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 D31.

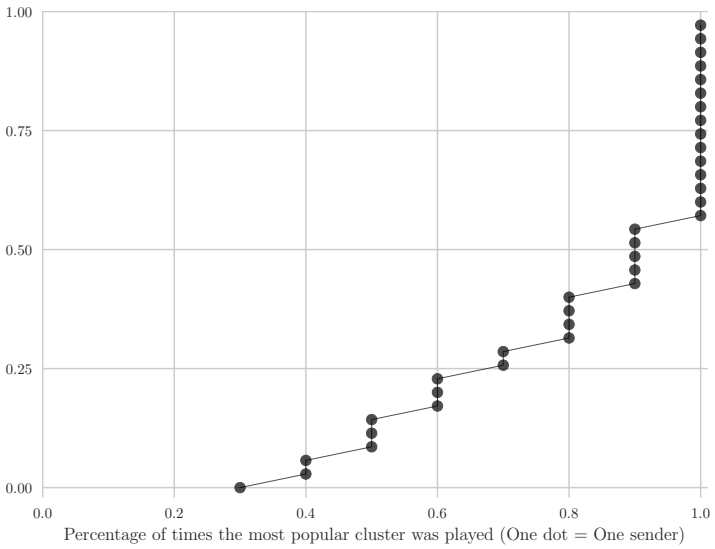


Figure D32: Persistence in senders' type

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.[42] 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.

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