NBER WORKING PAPER SERIES

RULES AND COMMITMENT IN COMMUNICATION:
AN EXPERIMENTAL ANALYSIS

Guillaume R. Fréchette
Alessandro Lizzeri
Jacopo Perego

Working Paper 26404
http://www.nber.org/papers/w26404

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2019, Revised August 2021

We thank Andreas Blume, Elliot Lipnowski, Salvatore Nunnari, Santiago Oliveros, Sara Shahanaghi, and Emmanuel Vespa for useful comments. We also thank the co-editor and three referees for very helpful comments. Fréchette and Lizzeri gratefully acknowledge financial support from the National Science Foundation via grant SES-1558857. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Rules and Commitment in Communication: an Experimental Analysis
Guillaume R. Fréchette, Alessandro Lizzeri, and Jacopo Perego
NBER Working Paper No. 26404
October 2019, Revised August 2021
JEL No. C92,D7,D8,D9

ABSTRACT

We study the role of commitment in communication and its interactions with rules, which determine whether information is verifiable. Our framework nests models of cheap talk, information disclosure, and Bayesian persuasion. It predicts that commitment has opposite effects on information transmission under the two alternative rules. We leverage these contrasting forces to experimentally establish that subjects react to commitment in line with the main qualitative implications of the theory. Quantitatively, not all subjects behave as predicted. We show that a form of commitment blindness leads some senders to overcommunicate when information is verifiable and undercommunicate when it is not. This generates an unpredicted gap in information transmission across the two rules, suggesting a novel role for verifiable information in practice.

Guillaume R. Fréchette
New York University
Department of Economics
19 West 4th Street
New York, NY 10012
guillaume.frechette@nyu.edu

Alessandro Lizzeri
Department of Economics
Princeton University
Julis Romo Rabinowitz Building
Princeton, NJ 08544
and NBER
lizzeri@princeton.edu

Jacopo Perego
Columbia University
3022 Broadway Uris 616
New York, NY 10027
jp3754@columbia.edu

1 Introduction

The goal of this paper is to experimentally study the effects of *rules* and *commitment* in communication. In our analysis, rules are restrictions on language that determine whether an agent can freely misreport what she knows or whether she can only use verifiable information. Commitment captures the extent to which the agent can communicate according to predetermined protocols. Any communication environment potentially can be affected by the degree of commitment and by the nature of the rules governing communication. For instance, models of cheap talk, information disclosure, and Bayesian persuasion differ from each other in ways that lead back to differences in rules and commitment. In many concrete applications, it is difficult to measure the exact degree of commitment available to an agent or the extent to which rules are enforced. Yet, rules and commitment do vary significantly in practice depending on the context and observables such as the frequency of communication. Thus, studying their effects on communication is a natural question.

We present a simple model of communication under *partial* commitment and consider two alternative rules: *verifiable* and *unverifiable* information. The focus on partial commitment is a key feature of our analysis: it allows us to nest many existing communication models under the same umbrella and experimentally test key qualitative predictions about the role of commitment in communication. The contrast between verifiable and unverifiable information further enriches our analysis, as the main comparative statics have opposite signs under these two alternative rules. Our main results indicate clear treatment effects in line with the main qualitative predictions of the theory. We also uncover important quantitative deviations from the theory. Specifically, we find that rules matter in unpredicted ways; we propose a systematic rationalization for these departures.

We consider a sender-receiver model with binary states and actions. The sender wants the receiver to choose a high action, whereas the receiver wishes to match the state. There are three stages. In the commitment stage, the sender publicly commits to an information structure, which is a map between states and messages. Under unverifiable information, the sender can freely misreport her private information. Under verifiable information, she can only conceal it. In the revision stage, the sender learns the state and can privately revise the chosen information structure. In the guessing stage, the receiver observes a message and chooses an action. The message is generated with probability ρ from the commitment stage and with the remaining probability from the revision stage. We view the probability ρ as capturing the sender's commitment power: the higher ρ is, the higher the probability that the sender will *not* be able to revise her strategy after learning the state and thus, the higher the extent to which she is committed to her initial communication. An observable prediction of the model is that variations in commitment power generate outcomes that are qualitatively different depending on the communication rule. For example, an increase in the sender's commitment power should *increase* the amount of information conveyed under unverifiable information, whereas it should *decrease* it under verifiable information. When the sender can fully commit, these two scenarios coincide and the information conveyed in equilibrium is independent of the communication rule. We exploit these predictions to experimentally test the role of commitment in communication.

This framework captures the flavors of a wide variety of models of communication, including models of cheap talk (Crawford and Sobel, 1982; Green and Stokey, 2007), disclosure (Grossman, 1981; Milgrom, 1981; Jovanovic, 1982; Okuno-Fujiwara et al., 1990), and Bayesian persuasion (Kamenica and Gentzkow, 2011). It helps organize our analysis in two ways. First, the comparison *across* models generates contrasting predictions that go to the heart of the strategic tension of communication under commitment. As we illustrate in the paper, these contrasts discipline which explanations can be used to rationalize potential departures from the theory. Second, the framework itself informs a parsimonious experimental design. In our treatments, we change two variables—the degree of commitment ρ and the verifiability of information—while leaving the underlying structure of the game unchanged.

We begin by establishing several patterns in the data that are consistent with the key qualitative predictions of the theory. Specifically, we present two main sets of findings. First, we show that, on average, both senders and receivers react to commitment. For senders, we exploit within-treatment variation to show that between the commitment and the revision stages, their average behavior changes in the direction predicted by the theory. When information is unverifiable, senders reveal more information in the commitment stage than in the revision stage. When information is verifiable, this ranking is reversed, as predicted by the theory. For receivers, we exploit across-treatment variation to show that, as commitment increases, they become more responsive to information from the commitment stage. These reactions are consistent with the fact that information conveyed in the commitment stage is more meaningful when the level of commitment is higher. For our second main finding, we test how increasing commitment power changes the amount of information conveyed by the senders. In line with the theory, we find that this amount increases with commitment in treatments with unverifiable information and decreases with commitment in treatments with verifiable information. Furthermore, we find that verifiability has the predicted effect of increasing the amount of information conveyed by senders. Overall, these strong treatment effects validate the qualitative implications of the theory, especially given the the contrasting implications of the theory depending on the verifiability of messages.

We then analyze the main quantitative deviations from the theory that we observe in the data. In treatments with low commitment, we replicate existing findings in the literature by showing that, relative to the predictions of the theory, senders undercommunicate when information is verifiable and overcommunicate when it is not.[1] However, we find that the opposite holds in treatments with high commitment: senders overcommunicate when information is verifiable and undercommunicate when it is not. These deviations create an *information gap* between verifiable and unverifiable treatments, which is particularly apparent in the limiting case of full commitment: empirically, the amount of information conveyed is higher in verifiable treatments than in unverifiable ones, even though in theory this amount should be the same. From a policy perspective, this information gap presents a novel justification for making it more difficult for senders to misreport their information.

We discuss the extent to which a model with boundedly rational agents may help explain these deviations. We note that a number of plausible biases that have been explored in prior work—such as lying-averse senders or non-Bayesian receivers— are insufficient to rationalize the observed deviations. We consider the possibility that a fraction of senders are *commitment blind*: they behave under commitment as if they had no commitment power whatsoever. That is, they are incapable of exploiting commitment to their advantage. In both stages, these senders choose a strategy that is optimal under no commitment. This bias has different implications depending on the communication rule and, in particular, could explain the observed information gap. To find evidence for commitment blindness, we look at treatments with partial commitment, where we can observe the behavior of the same sender in scenarios with and without commitment power. Our analysis reveals that there is a group of senders who behaves in ways that are consistent with commitment blindness. To evaluate whether this explanation is fully capable of accounting for the quantitative departures from theory, we estimate a structural model of Quantal Response Equilibrium (QRE). By clustering the observed senders' strategies in treatment-specific representative groups, we can capture the typical behavior of commitment-blind senders. For each treatment, we then simulate data from our estimated model and find that it can explain a considerable part of the gap observed in the data.

Related Literature. The role of commitment in communication is at the center of the recent literature on persuasion and information design (Kamenica, 2019; Bergemann and Morris,

[1] For cheap talk, see the survey by Blume et al. (2020). For information disclosure, see Jin et al. (2020) and references therein.

2019). To study the effects of commitment, we innovate by considering a versatile framework in which commitment can be varied experimentally. Recent theoretical contributions by Lipnowski et al. (2018) and Min (2017) more generally analyze the implications of partial commitment under unverifiable information.[2] In a framework with no commitment, Kartik (2009) studies changes in lying costs, bridging models of cheap talk and information disclosure.

Our paper relates to a large body of experimental literature on cheap talk, which has been recently surveyed by Blume et al. (2020). Models of cheap talk feature no commitment and unverifiable information and have been used to study a variety of phenomena, including lobbying (Austen-Smith, 1993; Battaglini, 2002) and the interaction between legislative committees and a legislature (Gilligan and Krehbiel, 1987, 1989). Dickhaut et al. (1995) was the first experimental paper to test the central prediction of Crawford and Sobel (1982) 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. By contrast, in the experiment, communication leads to additional trade, partly because receivers are too credulous. Blume et al. (1998) study a richer environment and compare behavior when messages have preassigned meanings with behavior when meanings emerge endogenously. Among other findings, they confirm that, as in Forsythe et al. (1999), receivers are gullible. Cai and Wang (2006) also vary preference alignment and find that senders overcommunicate relative to the predictions of the cheap talk model and that receivers are overly trusting.[3]

Our paper also relates to the literature on information disclosure. Disclosure models feature no commitment but verifiable information and have been used to study quality disclosure by a privately informed seller (e.g., Verrecchia, 1983; Dye, 1985; Galor, 1985). Milgrom (2008) and Dranove and Jin (2010) survey this literature. In contrast to experiments on cheap talk, experiments on the disclosure of verifiable information typically find that senders undercommunicate relative to the theoretical predictions. For instance, Jin et al. (2020) find that receivers are insufficiently skeptical when senders do not provide any information, which in turn leads senders to undercommunicate.[4] Jin et al. (2019) and de Clippel and Rozen (2020) find evidence for strategic obfuscation of verifiable evidence in settings with no and full commitment, respectively. Information unraveling has also been studied in the field. For instance, Mathios

[2] Perez-Richet and Skreta (2018) study a model of interim information manipulation under full commitment.

[3] See also Sánchez-Pagés and Vorsatz (2007), Wang et al. (2010), and Wilson and Vespa (2020).

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

(2000) and Jin and Leslie (2003) document the failures of information unraveling for food nutrition labels and hygiene grade cards in restaurants.

One of our treatments replicates the leading example in Kamenica and Gentzkow (2011) and is one of the first tests of Bayesian persuasion. This treatment features full commitment and unverifiable information.[5] Other papers have studied a similar treatment with different designs and goals. Aristidou et al. (2019) compare the design of information and monetary incentives. Their remarkably simple implementation imposes some aspects of the equilibrium behavior onto subjects' tasks. In their findings, senders are able to extract a higher rent from receivers when using information rather than monetary incentives. On average, senders' strategies are close to equilibrium—a result that is in line with one of our findings. Au and Li (2018) augment Bayesian persuasion with reciprocity and test their model in the laboratory. In their implementation, senders directly choose posteriors instead of information structures. This simplifies senders' tasks and eliminates the need for receivers to do Bayesian updating. Their results highlight interesting inconsistencies relative to the standard theory. Finally, Nguyen (2017) uses an intuitive interface for senders and allows them to choose among a small set of precompiled communication strategies. Overall, given receivers' behavior, a large fraction of senders behave optimally and their behavior involves partial information transmission.

2 Theoretical Framework

In this section, we present our theoretical framework and discuss its main predictions. The model achieves two goals. First, it captures settings in which the sender has only *partial* commitment power. Second, it highlights the contrast between *verifiable* and *unverifiable* information. These features generate a rich set of predictions that we then test experimentally.

2.1 Model

There are two players: a sender and a receiver. The sender privately observes a state and communicates with the receiver to influence her final decision, which affects everyone's payoff. More specifically, let $\theta \in \{\theta_L, \theta_H\}$ be the state and $\mu_0 \in (0, 1)$ denote the prior probability that the state is θ_H. The receiver chooses an action $a \in A = \{a_L, a_H\}$ and wishes to match her action

[5] In a different setting, the experimental literature on Cournot competition with endogenous timing also studies commitment in the lab. A player can choose to publicly commit to a production quantity, thus emerging as a Stackelberg leader and increasing her payoff. See, for instance, Huck and Müller (2000), Huck et al. (2001), and Morgan and Várdy (2004, 2013).

to the state. That is, her state-dependent payoff is

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

where the relative cost of the mistakes in the two states is parametrized by q. A rational receiver would choose action a_H whenever her posterior belief that the state is θ_H is larger than q. We call q the *persuasion threshold* and assume that $\mu_0 < q$. That is, with no communication, the receiver would choose a_L.[6] The sender earns a positive payoff only if she successfully persuades the receiver to take action a_H. Specifically, her payoff is $v(a) = 1$ if $a = a_H$, and $v(a) = 0$ otherwise.

Let an information structure be a map $\pi : \{\theta_H, \theta_L\} \to \Delta(M)$, where $M = \{\theta_H, \theta_L, n\}$ is an exogenously specified set of messages. Denote by Π^U the set of *all* such information structures and by Π the subset from which the sender can choose. The difference between Π and Π^U captures exogenous restrictions on the sender's strategies, which we call *communication rules*. If $\Pi = \Pi^U$, we say that information is *unverifiable*. In this case no restrictions are imposed on the sender strategies. Conversely, we say that information is *verifiable* if $\Pi = \Pi^V := \{\pi : \{\theta_H, \theta_L\} \to \Delta(M) \mid \pi(\theta_H|\theta_L) = \pi(\theta_L|\theta_H) = 0\}$. In this case, message $m = \theta$ can only be sent by type θ and, therefore, it represents a verifiable statement asserting that the state is indeed θ. In contrast, message $m = n$ can be sent by both types.

The game unfolds in three consecutive stages. In the *commitment stage*, before observing the state θ, the sender chooses $\pi_C \in \Pi$. In the *revision stage*, the sender privately observes θ and chooses $\pi_R \in \Pi$. Because π_R is chosen after observing θ, the sender has no commitment power in the revision stage. In the *guessing stage*, a message m is drawn with probability $\rho \in [0, 1]$ from $\pi_C(\cdot|\theta)$ and with probability $(1-\rho)$ from $\pi_R(\cdot|\theta)$. The receiver observes π_C and m, but does not observe either θ or π_R. She chooses an action $\sigma(\pi_C, m) \in \Delta(A)$.

We refer to ρ as the sender's *degree of commitment*. It captures the extent to which the sender is able to commit to her commitment-stage strategy π_C. For high values of ρ, the message m is more likely to be determined by strategy π_C, which is chosen before the sender has learned the state and it is publicly observed by the receiver. Conversely, for low values of ρ, the message m is more likely to be determined by the revision-stage strategy π_R, which is chosen after the sender has learned the state and it is not observed by the receiver. For this reason, we refer to π_C and π_R as the commitment and revision strategies, respectively.[7]

[6] When instead $\mu_0 \geq q$, revealing no information is optimal for the sender regardless of the degree of commitment and the verifiability scenario.

[7] Alternative but equivalent interpretations are possible. One can think of the sender as having the opportunity

In summary, our framework is characterized by three main variables, which are common knowledge among the players: (i) the communication rule, Π^U versus Π^V, (ii) the degree of commitment $\rho \in [0,1]$, and (iii) the persuasion threshold $q \in (\mu_0, 1)$. This framework is convenient as it allows us to span across notable communication models. When $\rho = 0$ and information is unverifiable, our model captures cheap-talk communication. When $\rho = 0$ and information is verifiable, our model captures a disclosure game with verifiable communication. Finally, when $\rho = 1$ and information is unverifiable, our model captures a Bayesian persuasion game.

As in many communication games, this framework features multiple Perfect Bayesian Equilibria (PBE), which are defined and discussed in Appendix A. In the paper, we impose a tie-breaking rule on the sender behavior that refines the set of equilibria. We assume that, in both the commitment and the revision stage, whenever two strategies lead to the same continuation payoff, the sender breaks ties in favor of the one with the highest probability of sending message $m = \theta_H$ conditional on state θ_H. The idea is that honesty is especially prominent when it is also convenient for the sender. In contrast, we do not impose any restriction on how the the sender should break ties conditional on θ_L. This tie-breaking rule is simple but powerful: it is sufficient to guarantee the uniqueness the equilibrium outcome. Moreover, it formalizes the tendency to use natural language that we see in the data. We refer the reader to Appendix A for further discussion about the refinement. In the rest of the paper, we will refer to PBE that satisfy this tie-breaking rule as *equilibria* without further qualification.

2.2 Main Predictions

In this section, we describe the main theoretical predictions that we later bring to the laboratory. To do so, we introduce two measures of the correlation between the and the action, denoted ϕ and ϕ^B. These measures quantify in different ways the extent to which the sender transmits information to the receiver.

Our first measure focuses on the joint behavior of sender and receiver. Let (π_C, π_R, σ) be a profile of strategies and define $\phi(\pi_C, \pi_R, \sigma) := \mathrm{Corr}_{(\pi_C,\pi_R,\sigma)}(\theta, a)$, the statistical correlation between the state θ and the action a that is induced by (π_C, π_R, σ). The correlation ϕ can be viewed as a measure of "information received," namely, the extent to which the receiver reacts

to revise her commitment strategy after learning the state, which occurs only with probability $1 - \rho$. Another interpretation 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.

to the information sent by the sender. It captures the informativeness of the outcome induced by the players' strategies.[8]

Our second measure focuses exclusively on the sender's behavior. Fix any information structure $\pi \in \Pi$. For example, this could be π_C, π_R, or $\rho\pi_C + (1-\rho)\pi_R$. Consider a hypothetical receiver with utility u and prior belief μ_0, who optimally responds to the message m drawn from π. That is, such a receiver chooses $\sigma^B(m) = a_H$ if $\mu_0\pi(m|\theta_H) \geq q(\mu_0\pi(m|\theta_H) + (1-\mu_0)\pi(m|\theta_L))$ and $\sigma^B(m) = a_L$ otherwise. Define $\phi^B(\pi) = \text{Corr}_{\pi,\sigma^B}(\theta, a)$, the statistical correlation between the state θ and the action a induced by (π, σ^B). We refer to ϕ^B as the "Bayesian" correlation. It can be viewed as a measure of "information sent," namely, the extent to which the sender conveyed useful information to a hypothetical Bayesian receiver. It captures the informativeness of the outcome induced by the sender's strategy and the behavior of such a receiver.

When (π_C, π_R, σ) is on the equilibrium path, $\phi(\pi_C, \pi_R, \sigma) = \phi^B(\rho\pi_C + (1-\rho)\pi_R)$, that is, the two measures coincide. However, distinguishing between ϕ and ϕ^B is useful for two reasons. First, in the experiment, receivers may of course fail to be Bayesian. In such a case, $\phi^B(\rho\pi_C + (1-\rho)\pi_R)$ will help us isolate the information sent by the sender net of the receivers' mistakes. Second, $\phi^B(\pi_C)$ and $\phi^B(\pi_R)$ allow us to quantify how much information is sent by the sender's behavior in the commitment and the revision stage.

We now characterize the equilibrium outcomes. We begin by fixing the degree of commitment ρ and the communication rule. We show uniqueness of the equilibrium correlation and compare the Bayesian correlation of the strategies in the commitment and revision stages. To this purpose, define $\underline{\rho} := \frac{q-\mu_0}{q(1-\mu_0)}$ and $\bar{\rho} := \frac{q(1-\mu_0)}{q(1-\mu_0)+(1-q)\mu_0}$.

Proposition 1. *Fix ρ and the communication rule. All equilibria induce the same correlation. In any equilibrium:*

- *If information is verifiable and $\bar{\rho} \leq \rho < 1$, then $\phi^B(\pi_C) < \phi^B(\pi_R)$.*
- *If information is unverifiable and $\underline{\rho} < \rho < 1$, then $\phi^B(\pi_C) > \phi^B(\pi_R)$.*

This result highlights a tension between the commitment and revision stages. This tension manifests itself in opposite ways under the two alternative communication rules, thus providing useful and testable predictions that we will exploit in our experimental analysis. The intuition for Proposition 1 is the following. Under both verifiable and unverifiable information, the sender would like to commit to persuading the receiver to choose the high action as often

[8] In Online Appendix D.2, we show that if $(\pi'_C, \pi'_R, \sigma')$ induces an outcome that is more informative than (π_C, π_R, σ) in the Blackwell sense, then $\phi(\pi'_C, \pi'_R, \sigma') \geq \phi(\pi_C, \pi_R, \sigma)$.

as possible. When ρ is sufficiently high, this implies that partial information revelation occurs in both verifiability scenarios. However, in the revision stage, the sender is unable to resist the temptation to undo her commitments and manipulate information in her favor. Under verifiable information, this opportunity implies that the strategy in the revision stage reveals the state ("unraveling"). Thus, $\phi^B(\pi_R) = 1$. Under unverifiable information, instead, it implies that the strategy in the revision stage is uninformative ("babbling"). Thus, $\phi^B(\pi_R) = 0$. A notable aspect of the behavior implied by Proposition 1 is that it features the opposite pattern depending on verifiability: in transitioning between commitment and revision stages information transmitted increases for verifiable information and it declines for unverifiable information. Interestingly, as we will show later in Table 2, in the commitment stage, the sender anticipates her future behavior in the revision stage and prepares accordingly: relative to the full-commitment scenario, she overcommunicates when information is unverifiable and undercommunicates when information is verifiable. These commitment strategies are an attempt to obtain final posteriors that are as close as possible to the full-commitment scenario. Overall, this result illustrates how changes in the rules can generate stark contrasts in the way senders react to commitment power.

Our next result describes how equilibrium informativeness changes with the degree of commitment and how this depends on the communication rule.

Proposition 2.

- *When information is verifiable, the equilibrium correlation ϕ weakly decreases in ρ. In particular, $\phi = 1$ if and only if $\rho < \bar{\rho}$.*
- *When information is unverifiable, equilibrium correlation ϕ weakly increases in ρ. In particular, $\phi = 0$ if and only if $\rho < \underline{\rho}$.*
- *When $\rho = 1$, equilibrium correlation ϕ is independent of the communication rules.*

This result illustrates that changes in commitment affect equilibrium correlation in starkly different ways depending on the communication rules. To understand this result, we first consider two extreme cases. When $\rho = 0$, the sender has no commitment power. When information is verifiable, unravelling occurs in equilibrium and, thus, the correlation is equal to 1. When information is unverifiable, babbling is only equilibrium and, thus, the correlation is equal to 0. As ρ increases, the revision stage becomes increasingly less likely, and the relevance of the commitment-stage strategy increases. This allows the sender to approach the optimal solution under full commitment, $\rho = 1$. When $\rho = 1$, the equilibrium correlation is independent of the rules of communication. To see this, note that when $\rho = 1$ and information is verifiable, the sender can replace the use of message θ_H with message n. By doing so, she can induce the same joint distribution over states and actions that is optimal under unverifiable information.

3 Experimental Design

In this section, we describe the laboratory implementation of our model, the main treatments that we conducted, and how we compute the correlations ϕ and ϕ^B from the data. We view our experimental design as a particularly useful framework to organize our analysis of commitment and communication rules. As we illustrate in the next sections, subject behavior in any given treatment is heterogeneous and challenging to evaluate on its own. In contrast, the comparison across treatments, along with the asymmetric nature of our predictions, goes to the heart of the strategic tension in our model.

3.1 Lab Implementation and Treatments

We begin by describing the implementation of the base game. A ball is drawn at random from an urn that contains three balls, one red and two blue. The message can be red, blue, or empty. The receiver earns \$2 if she correctly guesses the color of the ball. She earns nothing otherwise. The sender earns \$2 if the receiver guesses that the ball is red, irrespective of its color. Given this, the prior is $\mu_0 = 1/3$ and the persuasion threshold is $q = 1/2$. To present our results, we adopt the following notation to distinguish between states, messages, and guesses: the state θ is R or B; the message m is r, b, or n; and the receiver's guess a is *red* or *blue*.

The game has three stages.[9] In the commitment stage, the sender chooses an information structure. She does so via a simple graphical interface (see Online Appendix E.1). The sender selects $\pi_C(\cdot|\theta)$ by moving a slider, one for each state. The slider's bar is colored according to the conditional probabilities implied by the sender's choice. These probabilities are updated in real time in a table above the slider bar. In the revision stage, the sender learns the color of the ball θ. With the same interface as the one just described, she can revise the part of her strategy that concerns the *realized* state. We do not elicit the sender's choice for the state that did not realize. This design choice is a direct implementation of the game as we have described it. Moreover, it helps highlight the stark contrast between the commitment and revision stage.[10] In the guessing stage, the receiver observes the information structure chosen by the sender in

[9]In the laboratory, we referred to the these three stages with neutral labels: the *communication*, *update*, and *guessing* stage. In the remainder of the paper, we maintain instead the nomenclature introduced in Section 2.

[10]Of course, when $\rho = 1$, there is no revision stage and, therefore, it is not included in the design.

Table 1: Treatments Denominations

	Sender's Commitment Power		
Information	$\rho = 0.20$	$\rho = 0.80$	$\rho = 1$
Verifiable	$V20$	$V80$	$V100$
Unverifiable	$U20$	$U80$	$U100$

the commitment stage but not the one chosen in the revision stage. For this last stage, we use the strategy method, that is, we elicit the receiver's guess for each possible message she could receive. This allows the effective sample size to be increased considerably while keeping the receiver's task relatively simple.

We have a 2×3 factorial between-subject design, namely, each subject participates in a single treatment. Our experimental variables are the sender's commitment power ρ and the communication rules (verifiable versus unverifiable information). For each rule, we conducted three treatments with different degrees of commitment: $\rho \in \{0.20, 0.80, 1\}$. This gives us a total of six treatments, which constitute the bulk of our investigation. Treatments are denoted as illustrated in Table 1. 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.

Table 2 reports the equilibrium strategies for each treatment. Figure 1 reports the predicted equilibrium correlations. This set of treatments captures the key tensions of our model. First, treatments $V80$ and $U80$ reveal the tension between the commitment and the revision stage, as summarized by Proposition 1. This tension goes in opposite directions according to whether information is verifiable. Second, informativeness is increasing in ρ when information is unverifiable, while the opposite holds when information is verifiable. Third, treatments $U100$ and $V100$ are predicted to induce an identical outcome through senders' strategies that are substantially different. In the following sections, we will exploit these tensions to test the role of commitment and rules in communication.[11]

For each treatment, we conducted four sessions, for a total of 24 sessions. Each session included 12 to 24 subjects (16 on average), for a total of 384 subjects recruited from the NYU undergraduate population using *hroot* (Bock et al., 2014). At the beginning of each session, instructions were read aloud, and subjects were randomly assigned a fixed role: sender or re-

[11] In theory, ϕ^B is predicted to be 0 in $U20$ and 1 in $V20$, suggesting that the comparison with $U80$ and $V80$ is a one-sided statistical test. In practice, however, the observed ϕ^B is likely to be higher than 0 in $U20$ and lower than 1 in $V20$, as suggested by the prior experimental evidence on $U0$ and $V0$ (see Section 1). Thus, the comparative statics are falsifiable also because the comparison with $U80$ and $V80$ could display the wrong signs.

Table 2: Equilibrium Predictions

	Sender								Receiver		Correlation
	Commitment				*Revision*				*Guessing*		Coefficient
Treat.	State	Message			State	Message			Mes.	Guess	$\phi = \phi^B$
		r	b	n		r	b	n			
	R	1		0	R	1		0	r	*red*	
$V20$	B		x	$1-x$	B		y	$1-y$	b	*blue*	1
									n	*blue*	
	R	0		1	R	1		0	r	*red*	
$V80$	B		3/4	1/4	B		0	1	b	*blue*	0.57
									n	*red*	
	R	0		1					r	*red*	
$V100$	B		1/2	1/2					b	*blue*	1/2
									n	*red*	
	R	1	0	0	R	1	0	0	r	*blue*	
$U20$	B	x	x'	$1-x-x'$	B	y	y'	$1-y-y'$	b	*blue*	0
									n	*blue*	
	R	1	0	0	R	1	0	0	r	*red*	
$U80$	B	3/8	$5\alpha/8$	$5(1-\alpha)/8$	B	1	0	0	b	*blue*	1/2
									n	*blue*	
	R	1	0	0					r	*red*	
$U100$	B	1/2	$\alpha/2$	$(1-\alpha)/2$					b	*blue*	1/2
									n	*blue*	

In $V20$, $x, y \in [0, 1]$. In $U20$, $1 - \underline{\rho} < \rho x + (1-\rho)y$. In $U80$ and $U100$, $\alpha \in [0, 1]$.

ceiver. In each session, subjects played 25 paid rounds of the game described above, with random rematching between rounds. Thus, for each treatment, we observe an average of 800 unique sender-receiver interactions. At the end of every round, complete feedback was provided to both senders and receivers. Appendix E.2 contains the instructions for one of our treatments. 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, and ranged from \$12 to \$60. On average, sessions lasted 100 minutes. Our statistical analysis focuses on the last ten rounds to allow enough time for subjects to familiarize themselves with the interface and to learn the relevant strategic forces in the task they faced.[12]

3.2 Computing the Correlations

We quantify the information transmitted between sender and receiver by computing the correlations between state and action. State–action correlations have been extensively used in the experimental literature on communication.[13] To compute these correlations, we take advantage of our use of the strategy method in the communication and guessing stages to obtain significantly

[12]As can be seen in Appendix D.4, some aspects of behavior change over the course of the experiments.

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

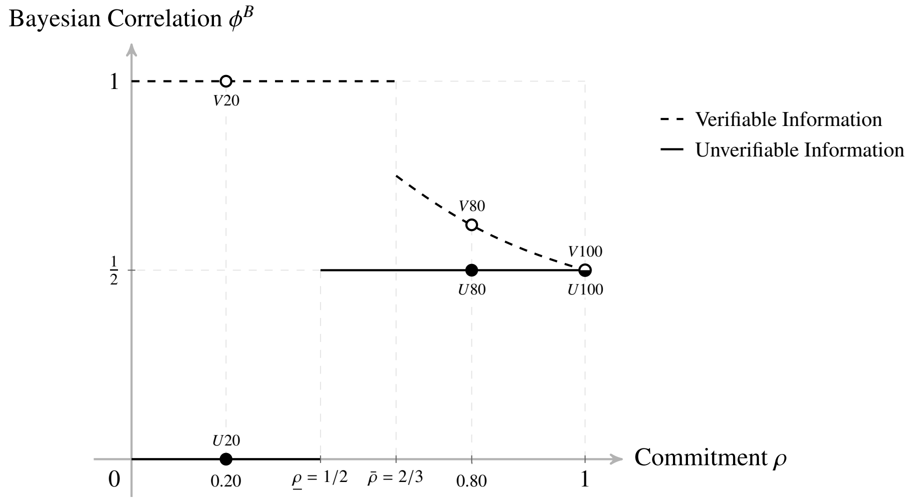


Figure 1: Predictions and Treatments.

more-precise measures of the correlation. However, in the revision stage, we observe only the sender's strategy conditional on the realized state θ. We circumvent this problem of missing data by imputing the session-specific average behavior of the senders in the revision stage.[14]

In Section 2.2, we distinguished between the correlation ϕ and the Bayesian correlation ϕ^B. The former uses the receiver's observed behavior and can be viewed as a measure of information received. The latter uses the behavior of a hypothetical Bayesian receiver, and can be viewed as a measure of information sent. Theoretically, there is no difference between ϕ and ϕ^B, as the receiver is assumed to be Bayesian in equilibrium. Empirically, however, ϕ and ϕ^B can differ because the former compounds the potential mistakes that receivers make when responding to the senders. For instance, if the sender truthfully discloses the state but the receiver does not listen, we would have that $\phi = 0$, despite a great deal of information being sent to the receiver.[15] As the central and the most novel aspect of our experiment is the behavior of senders, we will focus most of our attention on the Bayesian correlation ϕ^B.

3.3 Discussion of Design Choices

We briefly discuss our main design choices.

[14]This allows us to compute the correlations *for each* round, rather than taking averages *across* rounds. Through simulations, we verified that this leads to a substantial improvement in precision. Imputing session-specific averages seems a natural choice: due to the random rematching, receivers should hold comparable beliefs when facing a sender in the experiment. Our results are, however, robust to different imputation methods. For example, we can impute *subject*-specific averages and get essentially similar results. Also, it is important to note that the results for treatments with $\rho = 0.80$ (where we perform the imputation) are similar to those with $\rho = 1$ (where we do not need to use the imputation), suggesting the results are robust to our imputation method.

[15]The correlation ϕ can even be negative if the receiver were to grossly misinterpret the meaning of a message.

Treatments. It is instinctive to think of $\rho \in \{1/3, 2/3, 1\}$ as natural parametric choices. However, it is important to take into account the theoretical thresholds $\underline{\rho}$ and $\bar{\rho}$, defined in Section 2.2. In our experiment, $\mu_0 = ⅓$ and $q = ½$; thus, $\underline{\rho} = ½$ and $\bar{\rho} = ⅔$. We choose $\rho = 0.80$ to allow enough distance between the theoretical threshold $\bar{\rho}$, which is key for verifiable information, and the full-commitment benchmark. The choice of $\rho = 0.20$ ensures symmetry. In our treatments, we do not include the extreme case of $\rho = 0$ for two main reasons. First, this case is the only one for which there is experimental evidence already, both for verifiable and unverifiable information. Our main interest lies in treatments with partial and full commitment: these cases have not been tested in the laboratory and offer a unique opportunity to study the role of commitment in communication. Second, the equilibrium outcomes at $\rho = 0$ are identical to those at $\rho = 0.20$. In particular, the commitment power in treatments with $\rho = 0.20$ is so low that they could be seen as proxies for $U0$ and $V0$.[16]

Human Receivers. Senders' behavior is the central and more novel aspect of our experiment. Of course, senders' behavior depends on their expectation of how best to persuade receivers, which in turn depends on the receivers' observed behavior. One may think that there could be advantages to automating receivers' behavior to conform to the theory. We have three responses to this observation. First, we believe that senders' beliefs about how receivers interpret what message they see is central to understanding strategic communication. For instance, the main experimental finding in the literature on disclosure games, namely the failure of unraveling, would likely go undetected in a world with automated receivers. Second, the implementation of automated Bayesian receivers in the lab is far from trivial as it requires an explanation to senders of how the computer behaves. Failure to properly give this explanation defeats the potential purpose of introducing automated receivers. Moreover, it could generate demand effects as well as introduce additional complexity. Third, as we show in Section 5.1 and Appendix C, many of our receivers are non-Bayesian, but their behavior is systematic and is monotone in information, a property that is sufficient for our comparative static exercise.

Message n. From a theoretical perspective, the inclusion of message n in treatments with unverifiable information may seem redundant. However, in the experiment, it allows us to switch from unverifiable to verifiable information with minimal changes to our design. This increases our ability to compare results between different communication rules. It is perhaps reassuring to note that the majority of senders in treatments with unverifiable information employ a "natural" language—that is, message n is only marginally used.[17]

[16] We discuss this further in Online Appendix D.5.

[17] More specifically, the average total probability of message n, across all treatments with unverifiable information, is about 10%. In Appendix B.2, we compare $U100$ with a robustness treatment featuring a simpler message

Natural Language. Instead of using abstract labels for the messages, we label messages with colors that match the labels of the states—red and blue. In this way, messages can acquire a literal meaning. The focus of the paper is not on whether people understand how to coordinate on a language (Blume et al., 1998). Thus, we wished to remove one potential obstacle to communication that would have complicated the subjects' task and our analysis.

Additional Treatments. We conduct two robustness treatments, discussed in Appendix B. In our main treatments, payoffs are specified so that the persuasion threshold is $q = 1/2$. In an alternative payoff specification, we let $q = 3/4$. This allows us to test for changes in informativeness while keeping commitment and communication rules fixed. We also study a version of $U100$ with only two messages, r and b, and find that behavior in this robustness treatment is in line with $U100$, with slightly less noise.

Sender's Task. Our lab implementation of the sender's task is faithful to the nature of the game. In particular, the commitment stage involves a contingent choice, which is a random message for each state, while the revision stage is a single choice made after having learned the state. This could, in principle, make the commitment stage more complex for subjects than the revision stage. However, this differential complexity is embedded in the nature of commitment and not an artifact of the design.

Fixed Roles. Before the beginning of the experiment, subjects played two unpaid practice rounds in which they played the game from both the sender's and the receiver's perspective. Then, subjects were assigned to a fixed role—sender or receiver—and played that role for the duration of the experiment. Because the tasks that subjects faced in our experiment were nontrivial, we thought it would be important for them to gain relevant experience in their role.

Random Rematching. We chose to have random rematching of pairs of senders and receivers to simulate a one-shot interaction, while still allowing subjects to gain experience. Note, for instance, that experiments on duopoly games find that fixed pairing generates collusion, whereas random pairing does not (Huck et al., 2001).

4 Treatment Effects

In this section, we present the average treatment effects, which are in line with the main predictions of our theory. We discuss two main sets of results. In Section 4.1, guided by the predictions in Proposition 1 and Table 2, we look at how senders' behavior changes between the

space, $M = \{r, b\}$ instead of $M = \{r, b, n\}$. We find that subjects' behavior is highly comparable.

commitment and the revision stages as well as how receivers' responsiveness to information changes with commitment. In Section 4.2, we test Proposition 2 and analyze how the amount of information sent changes as we vary the level of commitment. Recall that a useful feature of our framework is that the predicted changes have opposite signs depending on verifiability.

We also document that subjects' behavior is highly heterogeneous. The treatment effects that we document are the result of the aggregation of different communication "styles." Although some subjects behave approximately as predicted by the theory, others either under- or over-react to commitment and rules. In Section 5, we will focus on these deviations to better understand their sources and implications.

4.1 Commitment and Subjects' Behavior

4.1.1 Senders

We begin by focusing on sender behavior. We explore the simplest and most direct evidence to test whether senders take advantage of commitment. By exploiting *within*-treatment variation in treatments $U80$ and $V80$, we observe how a sender's behavior changes between the commitment and the revision stages. Proposition 1 and Table 2 govern our predictions, which have opposite signs depending on whether the information is verifiable.[18]

Figure 2 displays the average difference in senders' strategies between the revision and the commitment stages in treatments $U80$ and $V80$. In the figure, a *positive* bar indicates a message that, conditional on the state, is sent more often in the revision stage. A *negative* bar indicates a message that is sent more often in the commitment stage.

Let us first consider treatment $U80$. Table 2 predicts that the sender should be more informative in the commitment stage than in the revision stage. In particular, when in the revision stage she learns that the state is B, she should replace message b with message r. That is, she should renege on her commitment to tell the truth. The results in the left panel of Figure 2 are very much in line with these predictions. Specifically, when the state is R, the equilibrium strategy is predicted not to change between the commitment and the revision stages. That is, all three bars should be of zero height. This is roughly what we observe in the data. Although statistically significant changes occur for r and b, they are tiny in magnitude.[19] Conversely, when

[18]We focus on $\rho = 0.80$ rather than $\rho = 0.20$ because, when $\rho > \bar{\rho} > \underline{\rho}$, the theory makes definite predictions about how senders' strategies and Bayesian correlations should change between stages.

[19] Unless noted otherwise, all statistical results allow for random effects at the subject level and are clustered at the session level. We include random effects to account for persistent heterogeneity across subjects; clustering is motivated by potential session effects (see Fréchette, 2012). Results for alternative specifications are reported

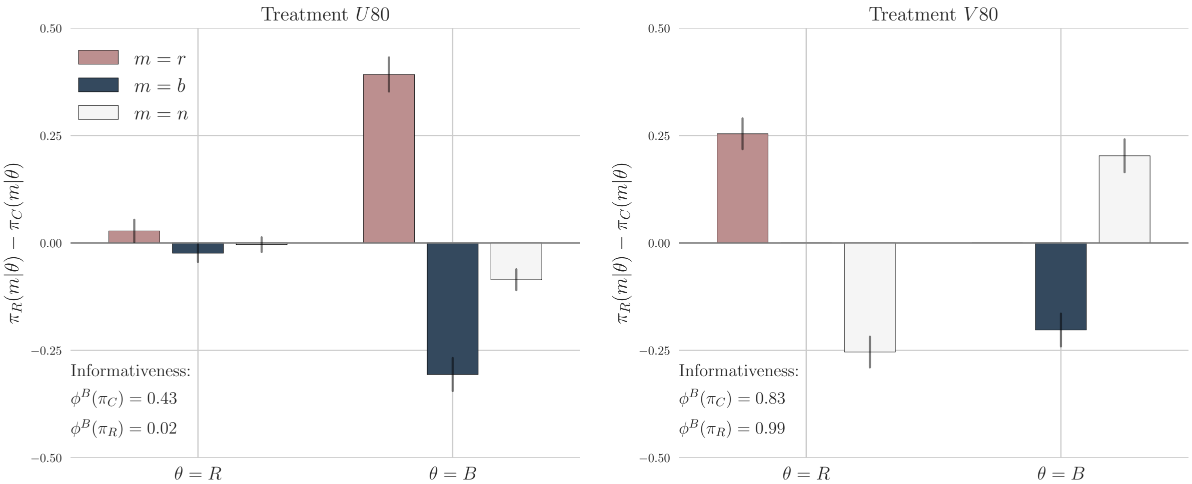


Figure 2: Sender's Strategy: Commitment vs. Revision, $\rho = 0.8$

the state is B, message r should replace b in the revision stage, whereas message n should not change. Again, qualitatively, this pattern is consistent with what we observe in the data. On average, senders increase the frequency of message r at the expenses of b ($p < 0.01$). Overall, as predicted by Proposition 1, the average informativeness of senders' strategies is significantly higher ($p < 0.01$) in the commitment stage—$\phi^B(\pi_C) = 0.43$—than in the revision stage—$\phi^B(\pi_R) = 0.02$.

We now turn to treatment $V80$ (right panel of Figure 2). Table 2 predicts the opposite type of behavior compared to $U80$: the sender should be less informative in the commitment stage than in the revision stage. In particular, when learning that the state is R, she should replace message n with message r, thus revealing the state. Furthermore, when learning that the state is B, she should replace message b with message n. These predicted changes are consistent with what we observe in the data. On average, when the state is R, senders entering the revision stage increase the likelihood of message r at the expense of message n. Instead, when the ball is B, they increase the likelihood of message n at the expense of message b. Both changes are significant at the 1% level. Overall, we find that the directions of the predicted changes are matched by the data as shown. Moreover, as predicted by Proposition 1, the average informativeness of senders' strategies is significantly lower ($p < 0.01$) in the commitment stage—$\phi^B(\pi_C) = 0.83$—than in the revision stage—$\phi^B(\pi_R) = 0.99$.[20]

in the appendix. We note that the findings in the alternative specifications suggest that session effects are not important in this setting. We have performed power calculations for key tests, for example, those associated with Figures 2 and 3, and established that at the estimated effect size, the power of our tests is well above the typical benchmark of 80%.

[20] We performed a similar analysis for $U20$ and $V20$ and found results that are roughly in line with those from treatments with $\rho = 0.80$. However, the interpretation of these results is more delicate because, since $\rho < \underline{\rho}$, these

From a quantitative point of view, unsurprisingly, sender average behavior falls short of exactly matching the equilibrium predictions. It is perhaps more interesting to note that most of the quantitative deviations come from behavior in the commitment stage. In contrast, average behavior in the revision stage is quite close to the theory. One possible explanation for these larger quantitative departures from the theory in the commitment stage is that this stage is more complex.[21] This distinction in the tendency of behavior to conform with theory in the two different stages has important consequences, as we discuss in Section 5.

In sum, the joint qualitative evidence arising from treatments $U80$ and $V80$ suggests that senders react to commitment and do so in ways that are consistent with the theory. One useful feature of considering different communication rules is that they generate opposing predictions within the same environment. On average, we see that senders exploit their commitment power to strategically hide good news (i.e., $m = n$ if $\theta = R$) when information is verifiable, and disclose bad news (i.e., $m = b$ if $\theta = B$) when information is unverifiable. Once in the revision stage, these commitments are no longer optimal, and indeed senders partially renege on them. We consistently observe the average informativeness of each stage changing as predicted.

4.1.2 Receivers

We now focus on receivers. Our goal is to evaluate the extent to which receivers respond to sender commitment and whether these responses are consistent with the theory. To explicitly test for this hypothesis we exploit *across*-treatment variations. We first introduce the idea of *interim* and *final* posteriors. Fix a commitment strategy π_C and a revision strategy π_R. An interim posterior is the belief that a Bayesian receiver would hold upon observing message if it was generated from the commitment strategy alone. That is, the interim posterior ignores the existence of the revision stage. The final posterior, instead, is the belief that such receiver would hold given that the message is generated from $\rho\pi_C + (1-\rho)\pi_R$. That is, the final posterior correctly takes into account the existence of the revision strategy π_R. Clearly, interim and final posteriors coincide when $\rho = 1$. More generally, given π_C and π_R, the higher the degree of commitment ρ, the closer the interim posterior is to the final one. We use this simple ob-

treatments lack clear-cut guidance from the theory for what concerns the sender's equilibrium strategy (see Table 2). Nonetheless, we still find that $\phi^B(\pi_C) = 0.48$ is higher than $\phi^B(\pi_R) = 0.00$ in $U20$ and that $\phi^B(\pi_C) = 0.88$ is lower than $\phi^B(\pi_R) = 0.94$ in $V20$.

[21] Evidence of this differential complexity may also be deduced from the fact that behavior is more heterogeneous in the commitment stage than in the revision stage in both $U80$ and $V80$ treatments. For example, in $U80$, the variance of commitment strategies is 0.43 while that of revision strategies is 0.28. The difference is significant at the 1% level. Similarly, for $V80$, the variance of commitment strategies is 0.45 while that of revision strategies is 0.23.

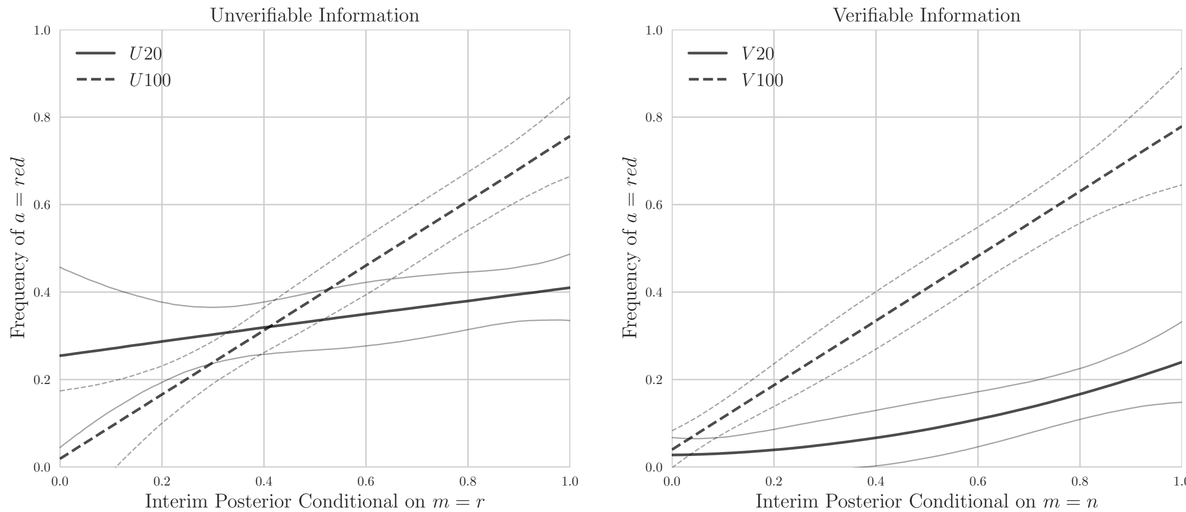


Figure 3: Receiver's Response to Persuasive Messages: $\rho = 0.2$ vs. $\rho = 1$

servation to test whether receivers respond to differing levels of commitment. 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 predictive of receivers' behavior; at low levels of commitment, they should not.[22]

This analysis is carried out in Figure 3. We look at how receivers' responsiveness to interim posteriors changes in treatments with low ($\rho = 0.20$) versus high ($\rho = 1$) commitment.[23] We plot polynomial fits of the average receiver's guess as a function of the interim posterior induced by the observed sender's π_C, the strategy from the commitment stage, and message m.

We begin by comparing treatments $U20$ and $U100$. Our focus is on message $m = r$. In $U20$, the interim posterior should have little or no impact on the receiver's guess because it is likely that message r did not come from the observed π_C. Therefore, the interim posterior is likely to be far from the final posterior. By contrast, in $U100$, the interim posterior should have a substantial positive effect on the probability that the receiver guesses *red* (Table 2). Indeed, interim and final posteriors coincide in this case. We report our results in the left panel of Figure 3. Consistent with the predictions, the estimated receivers' response is mostly flat in $U20$ and unresponsive to interim beliefs, whereas it is strictly increasing in $U100$.[24]

[22] An alternative approach to address the same question is to study how how receivers respond to identical commitment strategies π_C— as opposed to induced interim beliefs—in treatments with high versus low commitment. However, the space of commitment strategies is considerably larger and more complex than that of induced posteriors, which is $[0, 1]$.

[23] In Online Appendix D, Figure D17 performs the same exercise by comparing $\rho = 0.20$ and $\rho = 0.80$.

[24] The linearity in posteriors may be suggestive of *probability matching*. In Appendix C.1, we show that, instead, it results from aggregating the behavior of receivers who employ heterogeneous threshold strategies.

Similar—if not stronger—evidence is found when comparing $V20$ and $V100$ (right panel of Figure 3). By the nature of verifiable information, messages r and b induce trivial interim beliefs of either 1 or 0. For this reason, we focus on message n, which is the one requiring receivers to be sophisticated. We find that receivers' guessing behavior in $V20$ is quite flat in the interim posterior. In contrast, responsiveness is strong and positive for treatment $V100$.[25]

Overall, the joint evidence coming from Figure 3 suggests that, on average, receivers react to commitment in ways that are consistent with the theory. They correctly anticipate senders' incentives to renege on their commitments. As a consequence, receivers understand that messages inducing identical interim beliefs should be treated differently for different degrees of commitment.[26] Although this shows that receivers react to commitment, their behavior could still be far from Bayesian. Indeed, in line with a large body of experimental literature, Figure 3 suggests that this may be the case. We return to this point in Section 5 when we explore in detail the main quantitative deviations that we observe.

4.2 Commitment and Information Transmitted

The starkest prediction of our theory concerns how the correlation changes with the level of commitment under verifiable and unverifiable information. Proposition 2 predicts that equilibrium correlation should increase with commitment under unverifiable information, whereas it should decrease with commitment under verifiable information. To test this prediction, we compute the Bayesian correlation $\phi^B(\rho\pi_C + (1-\rho)\pi_R)$, which captures the amount of information sent. In Figure 4, we plot the cumulative distribution function (CDF) of the sender averages. That is, each dot represents the average Bayesian correlation induced by a sender in one of the treatments.

Two patterns emerge from this figure. First, when information is unverifiable (left panel), we observe a noticeable first-order stochastic *increase* in the information sent under $U100$ and $U80$ relative to $U20$. That is, the Bayesian correlation increases in commitment not only on average but at all percentiles of the distribution. Moreover, $U80$ and $U100$ are unranked, as predicted by the theory (Figure 1). Second, when information is verifiable (right panel), we observe a first-

[25] The probability that the receiver guesses *red* when the interim posterior is below $1/2$ does not differ statistically between $\rho = 0.2$ and $\rho = 1$, both for the case with unverifiable information (left panel) and verifiable information (right panel). Instead, for interim posteriors above $1/2$, we find a statistically significant difference in both cases ($p < 0.01$). Perhaps more importantly, the magnitude of the change—below and above $1/2$—is sizable: 56 versus 14 percentage points in the verifiable case, and 40 versus 6 percentage points in the unverifiable case.

[26] In Online Appendix D.6, we apply methods from Caplin and Martin (2021) to reach a similar conclusion. We find that receivers' behavior reveals that they are better informed in $U100$ rather than in $U20$.

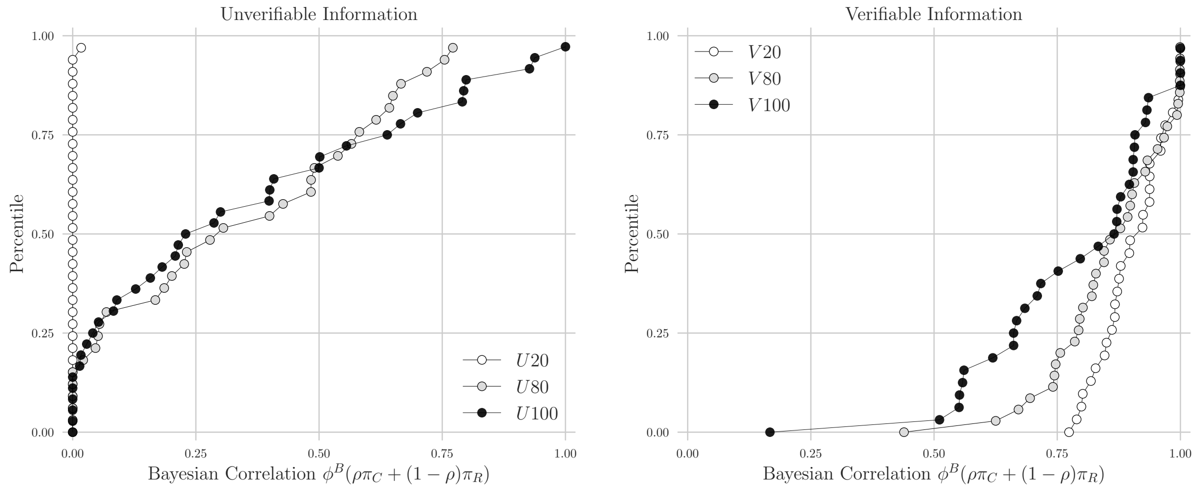


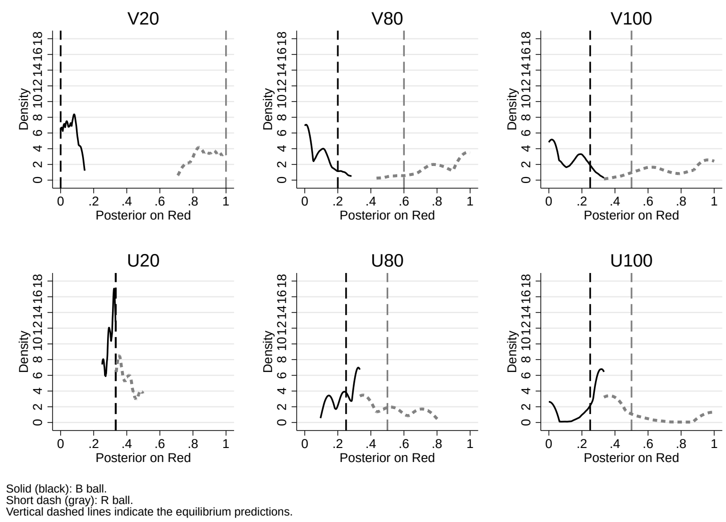
Figure 4: Cumulative Distribution of Sender-Average $\phi^B(\pi_C, \pi_R)$ by Treatment

order stochastic *decrease* in informativeness of $V100$ relative to $V20$. This change is relatively less pronounced in $V80$ relative to $V20$. Nonetheless, informativeness appears to decrease in commitment not just on average, but at all (or most, for $V80$) percentiles of the distribution. Again, this is consistent with the theory.

To provide further evidence on these comparative statics, we study an alternative measure of information sent. For every strategy profile (π_C, π_R), we compute $\psi^B = \mathbb{E}_m(\mu(m, \pi_C, \pi_R|\theta = R) - \mathbb{E}_m(\mu(m, \pi_C, \pi_R)|\theta = B)$, which is the divergence between the expected posterior conditional on the states.[27] The left panel of Figure 5 displays the kernel density estimates of the expected posteriors conditional on $\theta = R$ (in solid black) and on $\theta = B$ (dashed gray). The vertical dashed lines indicate the theoretical predictions. For instance, in $U100$, $\mathbb{E}_m(\mu(m, \pi_C, \pi_R|\theta = R) = 1/2$ because in equilibrium message r is sent with probability 1 and induces a posterior of 1/2. Instead, $\mathbb{E}_m(\mu(m, \pi_C, \pi_R|\theta = B) = 1/4$, because in equilibrium messages r and b are sent with 50% probability and induce posteriors of $1/2$ and 0, respectively.

In Figure 5, we see a sizable shift of the kernel distributions in the direction predicted by the theory, for both verifiable and unverifiable information. When commitment rises from $U20$ to $U100$, the two distributions become more spread out. In contrast, when commitment falls from $V20$ to $V100$, the posteriors move closer, as predicted by theory. These shifts are quantified in the right panel of Figure 5, which reports the average ψ^B as well as the average expected posteriors. The table shows that the data move in the right direction for both verifiable and unverifiable treatments, but that the mean difference is much closer to the theoretical predictions in the case of the unverifiable treatments than in the case of verifiable treatments.

[27] In Online Appendix D.2, we show that ψ^B is proportional to the posterior variance and, like ϕ^B, it is a completion of the Blackwell order on the senders' strategies.



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

Figure 5: On the left: Kernel Density of Expected Posterior Conditional on State. On the right: Average Differences in Expected Posteriors Conditional on State (theoretical values in parentheses)

Overall, the findings from Figure 4 and Figure 5 validate the contrasting comparative statics of Proposition 2. The theory is consistent with the main qualitative features of how senders' behavior changes with commitment and rules. Under verifiable information, senders use commitment to decrease the total amount of information they convey to receivers. Under unverifiable information, senders use commitment to increase the total amount of information they convey to the receivers. This contrasting use of commitment that we observe in the data suggests that, on average, sender behavior is consistent with the main strategic tension that underlies our model.

5 Understanding Departures from Theory

In the previous section, we showed evidence of treatment effects that match the main *qualitative* predictions of the model. Qualitatively, senders and receivers react to variations in commitment in the predicted ways. These treatment effects, however, hide substantial heterogeneity at the subject level which generates *quantitative* deviations from the theory. In this section, we document and offer an explanation for these deviations.

We begin by looking at the average Bayesian correlation by treatment. Table 3 reports the predicted Bayesian correlations (left panel) and the observed ones (right panel), averaged across sessions and subjects. These correlations move in the predicted direction as commit-

Table 3: Average Bayesian Correlations ϕ^B

ϕ^B – Theoretical Predictions

	Degree of Commitment ρ		
	$\rho = 0.2$	$\rho = 0.8$	$\rho = 1$
Verifiable	1	0.57	0.50
Unverifiable	0	0.50	0.50

ϕ^B – Observed

	Degree of Commitment (ρ)				
	$\rho = 0.2$		$\rho = 0.8$		$\rho = 1$
Verifiable	0.90	$\approx$	0.85	$>$	0.77
	$\vee$		$\vee$		$\vee$
Unverifiable	0.00	$<$	0.32	$\approx$	0.34

Notes: Symbol “>” indicates $p < 0.01$. Green symbol: as predicted. Red symbol: not as predicted.

ment changes. Moreover, in treatments with partial commitment, we note that more information is conveyed by the senders under verifiable information than under unverifiable information, in line with Section 4.2 and with our theory. However, Table 3 also highlights important quantitative deviations.

For each communication rule, the observed changes are more muted relative to the theoretical predictions. In the case of unverifiable information for example, the observed increase in ϕ^B from $U20$ to $U100$ is only 68% of the change predicted by the theory. In the case of verifiable information, the theory predicts that, moving from $V20$ to $V100$, we should observe a drop of 0.50 in the Bayesian correlation. Instead, in the data the corresponding reduction is only 0.13, or 26% of the predicted change. In particular, we find that, when commitment is high, senders tend to overcommunicate in treatments with verifiable information and undercommunicate in treatments with unverifiable information.

As a consequence of this phenomenon, verifiability affects the amount of information conveyed even when the theory predicts it should not. Most notably, Proposition 2 predicts that treatments $V100$ and $U100$ should generate identical Bayesian correlations. Instead, the observed Bayesian correlations are 0.77 and 0.34, respectively. This gap (significant at $p < 0.01$) represents a remarkable deviation from the theory. Furthermore, by comparing the black lines on the left and right panels of Figure 4, we can see that there is a gap at all percentiles of the distribution of ϕ^B, not just on average. More generally, in all treatments with high commitment, thus including $V80$ and $U80$, we observe that the Bayesian correlation is higher than predicted when information is verifiable whereas it is lower than predicted when information is unverifiable. We refer to these quantitative departures from the theory as the *information gap*.

In principle, this information gap could be due to anomalous behavior on the part of receivers, or of senders, or both. In Section 5.1, we explore receiver behavior and argue that, despite there being some observed departures from the Bayesian benchmark, it is unlikely that receivers are primarily responsible for this gap. In Section 5.2, we turn our attention to sender behavior. We show evidence of a behavioral bias that could explain the information gap. We call this bias *commitment blindness* and show that indeed it generates contrasting effects on the information transmitted depending on the communication rule and that it is therefore capable of generating the information gap. Finally, in Section 5.3, we estimate a structural model that accounts for such heterogeneity in sender behavior and show that it is capable of replicating in large part the observed deviations.

5.1 Can Receiver Behavior Explain the Information Gap?

Although Section 4.1.2 illustrates that receivers do react to commitment, a large body of experimental literature suggests that their behavior is likely to be non-Bayesian.[28] In Appendix C, we take a detailed look at receivers' behavior. Our analysis reveals that receiver behavior is indeed non-Bayesian. Yet, it is quite systematic. For example, most receivers behave in a way that is highly consistent with a "threshold" strategy: they guess *red* if the posterior is higher than some receiver-specific threshold. However, our analysis suggests that receiver behavior is unlikely to be the main explanation for the information gap. We discuss three main reasons for this conclusion.

First, we note that this gap cannot be *directly* determined by receivers' non-Bayesian behavior. Indeed, we expressed these gaps in terms of ϕ^B, the Bayesian correlation coefficient. By construction, this measure is immune to receivers' mistakes, as explained in Section 3.[29]

Second, we consider the possibility that the information gap could be *indirectly* generated by receivers, through the influence they exert on senders' strategies. For example, suppose that receivers are inherently skeptical of message n and respond to it by guessing *blue* regardless of the posterior (e.g., as in Jin et al., 2020). In treatments with unverifiable information, such a bias would have negligible consequences on senders' behavior: message n can be avoided in equilibrium and indeed is not used often in the data. In contrast, in treatments with verifiable information, message n plays a key role in the nature of the game. In the presence of such a bias against message n, the sender's optimal strategy in $V100$ must be fully informative, thus contributing to the information gap. However, we do not see evidence of such a bias among receivers in our data. Data show that receivers respond in similar ways to message n in treatments with verifiable information and r in treatments with unverifiable information.

[28] See, e.g., Charness and Levin (2005) and Holt (2007, Chapter 30) for an overview of such literature.

[29] When we explicitly include receivers' behavior—that is, when we compute ϕ instead of ϕ^B—we find informativeness gaps of similar magnitudes. In particular, we find that ϕ is 0.22 and 0.68 for treatments $U100$ and $V100$, respectively. Similarly, we find that ϕ is 0.19 and 0.78 for treatments $U80$ and $V80$, respectively.

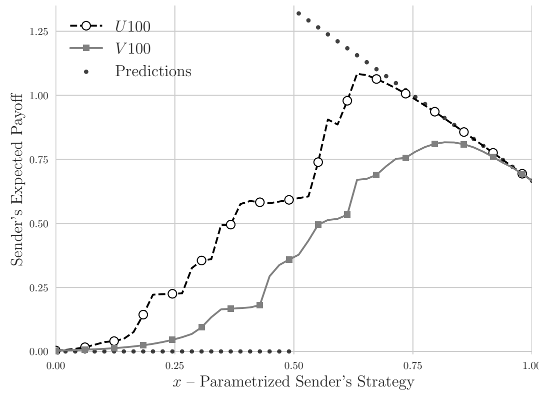


Figure 6: Sender's Empirical Expected Payoff

This can be seen in Figure 3. The dashed lines report receivers' responsiveness to message r in $U100$ (left panel) and n in $V100$ (right panel), controlling for their induced posterior. Receivers' responsiveness does not appear to be significantly different in the two cases, and they are highly responsive to the induced posterior for both messages.

Third, the information gap could be *indirectly* generated by receivers' non-Bayesian behavior, but in ways that are more complicated than our previous argument. To address this point, we estimate a simple model of receivers' behavior and then compute the sender's empirical best response. To be concise, we focus attention on treatments with full commitment. In Figure 6, we report the expected payoff that a sender would earn by playing various commitment strategies π_C when facing a *typical* receiver in our sample. For each treatment, we first fit a probit model to estimate the probability that $a = red$ given the message m, its induced posterior, and a subject fixed effect. Second, we use the estimated model to compute the expected payoff that a sender would earn when choosing various commitment strategies π_C. More specifically, we define a class of information structures parametrized by $x \in [0, 1]$. This class is rich enough to approximate most of the observed strategies, including the equilibrium strategies for these treatments. In particular, for $U100$, we consider strategies such that $\pi_C(r|R) = 1$ and $\pi_C(b|B) = 1 - \pi_C(r|B) = x$. For $V100$, we consider strategies such that $\pi_C(n|R) = 1$ and $\pi_C(b|B) = x$. In both $U100$ and $V100$, π_C is the equilibrium strategy when $x = 1/2$ (Table 2); it is uninformative when $x = 0$; it is fully informative when $x = 1$. More generally, $\phi^B(\pi_C)$ is weakly increasing in x.

Figure 6 shows that receiver behavior leads to a payoff function for the sender that is flatter than it would be if all receivers were fully Bayesian. Moreover, for both treatments, the sender's best response to the receivers' behavior requires $x > 1/2$. This is intuitive: $x = 1/2$ is a knife-edge condition that leaves a Bayesian receiver just indifferent. Although receivers do not conform with the Bayesian paradigm, the vast majority of them are more likely to guess *red* following a message that carries evidence that favors state R. This monotone responsiveness in induced beliefs is a milder rationality requirement than Bayesianism, and it has been documented in other experiments (see Camerer, 1998, for a discussion). Importantly, as shown by Figure 6, the extent of monotonicity displayed in our experiment is sufficient to confirm a key insight from models of communication under commitment, namely the fact that the best-response involves some degree of strategic obfuscation.[30] This analysis allows us to conclude that, given behavior by receivers in our data, an uninformative π_C is worse than a fully informative π_C, which is in turn worse than commitment to mixing. The finding that senders' empirical expected payoff is nonmonotone in the amount of information conveyed to the receiver is consistent with a key force of the theory.

Returning to the information gap, Figure 6 shows that receiver behavior alone appears insufficient to explain the large gaps in ϕ^B that we documented in Table 3. If senders were best-responding to the typical receiver behavior, we would observe $\phi^B(\pi_C) = 0.60$ in treatment $U100$ and $\phi^B(\pi_C) = 0.75$ in treatment $V100$. This explanation is, therefore, unsatisfactory on two levels. First, it captures only a small fraction (35%) of the observed gap. Second, the empirical best response for $U100$ would lead to an *increase* in informativeness, rather that the decrease that we observe in $U100$.

Overall, the three points above suggest that receivers' nonequilibrium behavior is insufficient to explain the informativeness gap. As we show in the remainder of the section, a bias in sender behavior is likely to be the primary driver of these observed deviations.

5.2 Commitment Blindness

In this section, we introduce a simple bias in senders' behavior that can explain a large part of the informativeness gap. We begin by noting that senders employ heterogeneous communication "styles," as already illustrated in Figure 4. Understanding the sources of this heterogeneity is key to explaining the information gap.

To this end, we introduce the notion of *commitment blindness*. We say that a sender is commitment blind if she behaves under commitment as if she had no commitment power at all. More specifically, her commitment strategy is the *equilibrium* strategy of a hypothetical game in which there is no commitment so that there is only a revision stage (i.e., a game with

[30] Relatedly, de Clippel and Zhang (2020) explore the relative robustness of the Bayesian persuasion model if the receiver is non-Bayesian.

$\rho = 0$). Commitment blindness has very different implications on the Bayesian correlation ϕ^B, depending on the communication rule. Specifically, when information is unverifiable, $\rho = 0$ is equivalent to a cheap talk game and any equilibrium strategy involves babbling. Such a strategy is *uninformative* ($\phi^B = 0$). If instead information is verifiable, $\rho = 0$ is equivalent to an information disclosure game and the equilibrium strategy involves unraveling; hence, it is *fully informative* ($\phi^B = 1$). If some of the senders were indeed commitment blind, their behavior could contribute to the information gap that we have documented. Indeed, relative to the theoretical prediction with fully rational senders, this bias tends to increase ϕ^B in treatments with verifiable information and to decrease it in treatments with unverifiable information.

Note that commitment blindness is different from lying aversion and leads to different implications. To see this, consider a sender who is fully averse to lying, regardless of her commitment power. First, when information is unverifiable, such a sender would play highly informative strategies in the commitment stage, in contrast with commitment blindness. Second, her behavior would increase the observed ϕ^B rather than decreasing it and, thus, it cannot generate the information gap that we observe in the data.

We exploit our experimental design to detect the presence of senders who display commitment blindness. This evaluation can only be done in treatments with partial commitment. Indeed, one needs to observe how the *same* sender behaves in two different commitment scenarios: with and without commitment power. We focus our attention on treatments $U80$ and $V80$ and compare how sender behavior changes between the commitment and the revision stages.[31] We seek to identify senders who (i) play the *same strategy* in both commitment and revision stages, and, (ii) play the *equilibrium strategy* in the revision stage as defined in Table 2.

In contrast to the previous discussion, we now want to understand more deeply the nature of heterogeneity in senders' behavior, and we do so by considering fully disaggregated data.[32] Our goal is to identify the most representative strategies (π_C, π_R) played in the treatments under consideration. Such an analysis presents a technical challenge, as senders' strategies are complex and high-dimensional objects. To organize the observed strategies, we use a standard machine learning algorithm, k-means, to cluster strategies into four representative groups.[33] We

[31] We focus on $\rho = 0.80$ instead of $\rho = 0.20$ because the information gap is a departure from the theory only for treatments with high commitment.

[32] It is then natural for the purposes of this section to impute revision-stage missing data using averages at the subject level rather than at session level (see Section 3.2).

[33] The k-means algorithm (see, MacQueen, 1967; Hastie et al., 2009; Murphy, 2012) is a commonly used method to cluster data. The procedure finds k clusters and their "centers" to minimize the total within-cluster variance. We set $k = 4$ and input an 8-dimensional vector of entries: $(\pi_C(m|\theta), \pi_R(m|\theta))$ for $m \in \{r, b\}$ and $\theta \in \{R, B\}$. Our conclusions from this exercise are robust to the choice of a different number of clusters.

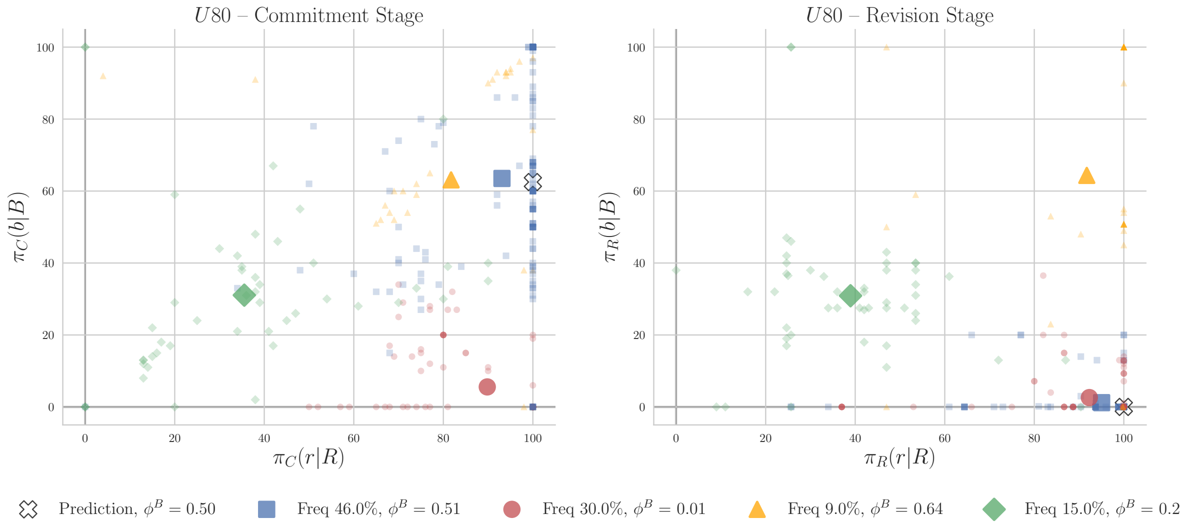


Figure 7: Treatment $U80$ – Clustering of Senders' Strategies

cluster the strategies by treatment and report the results in Figures 7 and 8 for treatments $U80$ and $V80$, respectively. To visualize all the data, we plot the clustered strategies onto two separate panels, one for π_C and one for π_R. The representative strategies that emerge from the algorithm are indicated with larger markers. Note that strategies (π_C, π_R) that appear similar in the commitment (revision) stage may belong to different clusters because they differ in the revision (commitment) stage.[34]

We begin our analysis with treatment $U80$, that is, Figure 7. The strategies indicated by red circles are those compatible with commitment blindness. The representative strategy consists of sending message r regardless of the state, in both the commitment and the revision stage. This strategy coincides with equilibrium behavior in the revision stage (Table 2). As expected, this strategy is almost completely uninformative ($\phi^B = 0.01$). This strategy is also quite common: 30% of the observed strategies are of this kind. We now discuss the remaining clusters of Figure 7. The strategies indicated by blue squares are compatible with equilibrium behavior and are the most prevalent ones. These strategies drive most of the treatment effects documented in Section 4. Note that their induced Bayesian correlation, $\phi^B = 0.51$, is remarkably close to the equilibrium prediction of 0.50. Strategies indicated by yellow triangles are consistent with a weak form of lying aversion and are not prevalent in our data. Finally, strategies marked by green diamonds belong to a residual cluster that cannot be grouped in any of the categories above. We interpret these residual strategies as noise.

We now turn to the analysis of sender behavior in treatment $V80$ (Figure 8). Again, strate-

[34]We present data at the observation level, but these clusters capture persistent sender types, with a typical sender playing in the same cluster more than 80% of the time.

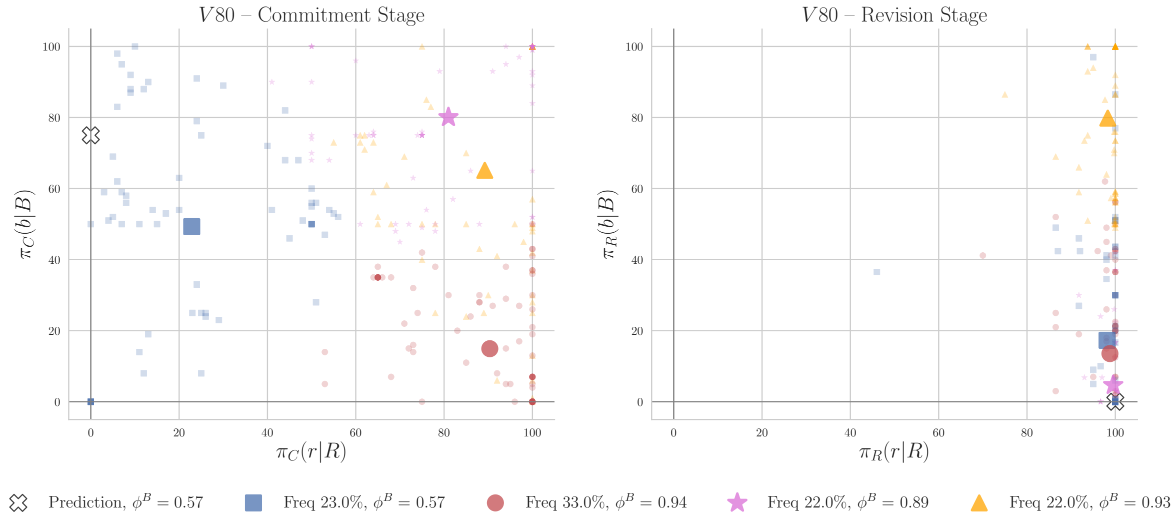


Figure 8: Treatment $V80$ – Clustering of Senders' Strategies

gies indicated by red circles are those compatible with commitment blindness. The representative strategy consists of sending message r given R, and n given B, in both the commitment and the revision stages. This coincides with equilibrium behavior in the revision stage (Table 2). In contrast to $U80$, commitment-blind strategies are highly informative ($\phi^B = 0.94$). In terms of prevalence, 33% of the observed strategies are of this kind. We now discuss the remaining clusters of Figure 8. Strategies indicated by blue squares are consistent with equilibrium behavior. They react to commitment and induce a Bayesian correlation of $\phi^B = 0.57$, which coincides with the equilibrium prediction. Strategies indicated by purple stars also react to commitment and play the equilibrium strategy in the revision stage but fail to conceal information in the commitment stage. As a result, they induce higher than optimal levels of Bayesian correlation, that is, $\phi^B = 0.89$. Together, these last two clusters we discussed represent 45% of the data and drive the treatment effects documented in Section 4. Finally, strategies indicated by yellow triangles are consistent with a weak form of lying aversion and induce high Bayesian correlation, that is, $\phi^B = 0.93$.

In sum, we have documented the existence of a behavioral type that is consistent with commitment blindness. Such behavior has opposite implications depending on the communication rule. Under unverifiable information, the behavior of these senders decreases the average Bayesian correlation ϕ^B. Under verifiable information, this behavior increases ϕ^B. Thus, a single behavioral bias can explain the main departures from equilibrium documented in Table 3.

We conclude our discussion by emphasizing some caveats. We have documented a specific behavioral bias without offering an explanation for *why* some subjects are biased in such way. For instance, we cannot distinguish whether commitment blindness is rooted in the fact that some senders misunderstand the meaning or the technology of commitment—which is what we emphasized above—or if they hold wrong beliefs about receivers' behavior. Likewise, it could be that these subjects engage in a specific form of backward anchoring, where behavior in the commitment stage imitates that in the revision stage.[35] Finally, the differential complexity of the commitment and the revision stages may contribute to the prevalence of this bias. Our experimental design is not set up to discriminate between these alternatives but we believe that it would be fruitful to explore them in future work.

5.3 QRE: Quantifying Departures From Equilibrium

In this final part of the section, our goal is to quantitatively reproduce the information gap with a structural model. The model we estimate has two components. First, it accounts for the heterogeneity in senders' behavior that we documented in Section 5.2. Second, it accounts for noisy players' behavior by using a quantal-response equilibrium (QRE) model (see, e.g., Goeree et al., 2016). With the estimated model, we compute the implied correlations and show that they account for about 70% to 80% of the observed information gap.

In a QRE, players make mistakes when responding to their beliefs, which however correctly account for the mistakes that other players make. Two technical challenges make the estimation of our structural model nontrivial. First, senders choose among a continuum of high-dimensional strategies and, second, our game is multi-stage and has incomplete information. We address the first challenge by using the same k-means algorithm that we discussed in Section 5.2. We address the second challenge by using the methodology in Bajari and Hortacsu (2005). In the following paragraphs, we explain these two points in more detail. For simplicity, our analysis focuses on treatments $U100$ and $V100$. Although a similar analysis could be performed under partial commitment, the focus on full commitment significantly simplifies our estimations. Moreover, the observed information gap in these treatments is maximal, so it is the one that is more interesting to explain.

Discretization of Senders' Strategies. To estimate QRE, we first need to discretize senders' strategy space into k representative strategies. Because our focus is on treatments with full commitment, the relevant strategy space is only comprised of the commitment-stage strategy, $\pi_C \in \Pi$. To find the k representative strategies, we use the same the same k-means algorithm

[35] We examined the class of "inertial" strategies, which are defined as those for which the Euclidean distance between π_C and π_R is especially small. We find that commitment-blind strategies make up an overwhelming majority of these inertial strategies.

Table 4: QRE-Implied Correlations

	Bayesian Correlation ϕ^B		Correlation ϕ	
Treatment	QRE-Implied	Observed	QRE-Implied	Observed
$V100$	0.72	0.77	0.64	0.68
$U100$	0.41	0.34	0.26	0.22

discussed in the previous section, and we set $k = 4$.[36] Importantly, we compute the representative strategies separately for each treatment. This allows us to capture the very different ways in which senders play in treatments with verifiable and unverifiable information, as shown in Section 5.2. In particular, it allows us to capture the different implications of commitment blindness for these two treatments.

Multi-Stage QRE. We assume that each player has a treatment-specific type, $\lambda_S \geq 0$ for the sender and $\lambda_R \geq 0$ for the receiver, and that these are common knowledge between players. We begin by describing the receiver behavior. Denote by $U(a \,|\, \pi_C, m) = \mu(m, \pi_C)\mathbb{1}(a = red) + (1 - \mu(m, \pi_C))\mathbb{1}(a = blue)$ the receiver's expected payoff from choosing action a conditional on observing the sender's commitment strategy π_C and the realization of message m. The QRE model assumes that a receiver of type λ_R guesses *red* with probability:

$$\mathbb{P}(red \,|\, \pi_C, m, \lambda_R) = \frac{e^{\lambda_R U(red \,|\, \pi_C, m)}}{e^{\lambda_R U(red \,|\, \pi_C, m)} + e^{\lambda_R U(blue \,|\, \pi_C, m)}}.$$

That is, the receiver can make mistakes, that is, choose a suboptimal action, and the probability of doing so decreases in λ_R as well as in the utility difference between the actions. We now turn to the sender behavior. Given the behavior of the receiver, the sender's expected utility from choosing strategy π_C is $V(\pi_C|\lambda_R) = \sum_{\theta,m} \mu_0(\theta)\pi_C(m|\theta)\mathbb{P}(red|\pi_C, m, \lambda_R)$. That is, the sender takes the receiver's mistakes into account when computing her expected payoff from playing a certain strategy. The probability that a sender of type $\lambda_S \geq 0$ chooses π_C is then given by

$$\mathbb{Q}(\pi_C|\lambda_S, \lambda_R) = \frac{e^{\lambda_S V(\pi_C|\lambda_R)}}{\sum_{\pi_C \in \Pi_k} e^{\lambda_S V(\pi_C|\lambda_R)}},$$

where Π_k denotes the discretized set of sender strategies discussed in the previous paragraph.

In sum, the model is pinned down by three parameters: Π_k, which we compute via the k-means algorithm; and λ_S and λ_R, which we estimate via maximum likelihood. The parameters

[36] Figure D18, in Online Appendix D, reports k-means clusters for treatments $U100$ and $V100$. Our results in this section are robust to choosing a different k.

(λ_S, λ_R) capture the extent to which players best respond to their opponent's behavior. At one extreme, as $\lambda_i \to \infty$, the player in role i never makes a mistake. At the other extreme, when $\lambda_i = 0$, the player in role i randomizes uniformly across all available strategies.

Estimation. We now describe how we estimate λ_S and λ_R. Recall that in treatments $U100$ and $V100$, the receiver observes the strategy π_C chosen by the sender. Whether this strategy was chosen by mistake is irrelevant for the receiver, who simply responds to π_C and its realized message m as described above. In other words, the receiver faces a single-agent decision problem. Thus, we can estimate λ_R independently of λ_S. In contrast, the sender moves before the receiver and, thus, the estimated value of λ_S will depend on the true λ_R. We consistently estimate $V(\pi_C|\lambda_R)$ for each strategy $\pi_C \in \Pi_k$ by computing the empirical average of the sender's expected payoff across the strategies that belong in the same cluster as π_C (Bajari and Hortacsu, 2005). Using maximum likelihood, it is then straightforward to estimate $(\hat{\lambda}_S, \hat{\lambda}_R)$.[37]

Simulation. Given these estimates, we simulate a large dataset with 10^4 hypothetical sender-receiver interactions. Each interaction comprises of a random θ, a strategy π_C chosen according to $\mathbb{Q}$, a message m, and a final guess a chosen according to $\mathbb{P}$. With this dataset, we can compute the correlation $\hat{\phi}$ and Bayesian correlation $\hat{\phi}^B$.

In Table 4, we report both the QRE-implied correlations as well as the observed ones. The main conclusion from this table is that the structural model we estimated generates correlations that are remarkably close to those we observed. In particular, the model explains between 70% and 80% of the observed information gap. It is useful to point out that, in the procedure described above, we fit data in two separate steps. First, we use the data from each treatment to compute Π_k from the k-means algorithm. That is, the representative strategies of treatment $U100$ are allowed to differ from those for treatment $V100$. We do so because, as revealed by our analysis in Section 5.2, communication rules affect senders' play in a substantial and unpredicted way. Second, we use the data again to estimate treatment-specific types $(\hat{\lambda}_S, \hat{\lambda}_R)$. By doing so, we allow the model to account for the mistakes that senders and receivers make when choosing their strategies. Thanks to the combination of these ingredients, the model is able to generate correlations that fit the observed data and replicate to a large extent the information gap.

[37] When $k = 4$, we find that $(\hat{\lambda}_S, \hat{\lambda}_R) = (0.41, 1.68)$ for $U100$ and $(\hat{\lambda}_S, \hat{\lambda}_R) = (0.21, 1.28)$ for $V100$.

5.4 Alternative Approaches

We now briefly discuss whether other theories could in principle account for the information gap: level-k, other-regarding preferences, and lying aversion. Although behaviors compatible with these theories may be present in our data to some extent, we argue that they are not the most natural avenues to explore, as they either fail to account for some of the key deviations from rational behavior, or they would need to be enhanced relative to their standard specifications.

Let us begin by considering a simple level-k model, which is a useful way to model strategic uncertainty.[38] A key starting point in such a model is the specification of how level-0 players behave. Importantly, in treatments $U100$ and $V100$, receivers do not face strategic uncertainty. Rather, they observe the strategy that the sender played and which message realized from it. In other words, these receivers face a single-agent decision problem, and it is not clear how to model their level-0 behavior. This observation implies that there is not much scope for the interesting feedback that sometimes occurs in a level-k analysis: any departure from the theory would be determined by assumptions about level-0 senders. Regarding senders, we have already discussed in Section 5.3 the consequences of noisy behavior. Two alternative types of level-0 senders are (i) truthtellers or (ii) senders who always send the same message regardless of the state. The first alternative would lead to an increase in correlation, both for $U100$ and $V100$; the second would lead to a decrease in correlation, both for $U100$ and $V100$. Therefore, these alternatives would lead to a unidirectional change in correlations that would not help close the information gap.

Other-regarding preferences have been successfully used to understand important patterns in a variety of experiments (see Cooper and Kagel, 2016). However, the information gap entails departures that, in some cases, go in a direction that is opposite to the common prediction of such models—namely, away from equating players' payoffs. For instance, in $U100$, a commitment-blind sender plays an uninformative strategy and thus earns the lowest possible payoff (see Figure 6), while the receiver can secure an expected payoff of \$1.33 (or \$2 times $2/3$) by guessing blue. By playing the empirical best response, the sender would instead increase her payoffs away from zero, while also increasing the payoff for the receiver. This suggests that commitment-blind senders do not behave in a way that is compatible with the spirit of many models of other-regarding preferences. Of course, this literature is incredibly rich, and there may be additional and more-complex types of behaviors that could be useful to explore

[38] Crawford et al. (2013) reviews this literature. In cheap talk games, Cai and Wang (2006), Kawagoe and Takizawa (2009), and Wang et al. (2010) discuss level-k models.

in the future.

Finally, lying aversion has been studied in the context of cheap talk experiments (e.g., Gneezy, 2005; Sánchez-Pagés and Vorsatz, 2007; Hurkens and Kartik, 2009). Lying aversion is consistent with the fraction of subjects who always tell the truth, as discussed in Section 5.2. However, such behavior is markedly different from the behavior of a commitment-blind sender, especially in treatments with unverifiable information. More importantly, it leads to implications that are, in principle, different from the observed departures: as mentioned earlier, lying aversion contributes to inflating the correlation in treatments with unverifiable information, whereas the opposite happens in the data.

References

Aristidou, A., G. Coricelli, and A. Vostroknutov (2019): "Incentives or Persuasion? An Experimental Investigation," *Working Paper*.

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

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

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

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

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

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

Bergemann, D. and S. Morris (2019): "Information Design: A Unified Perspective," *Journal of Economic Literature*, 57(1), 44–95.

Blackwell, D. A. and M. A. Girshick (1979): *Theory of Games and Statistical Decisions*, Courier Corporation.

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

Blume, A., E. K. Lai, and W. Lim (2020): "Strategic Information Transmission: A Survey of Experiments and Theoretical Foundations," in *Handbook of Experimental Game Theory*, ed. by C. M. Capra, R. Croson, M. Rigdon, and T. Rosenblat, Edward Elgar Publishing.

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

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

Camerer, C. (1998): "Bounded Rationality in Individual Decision Making," *Experimental Economics*, 1, 163–183.

Cameron, A. C. and D. L. Miller (2015): “A Practitioner’s Guide to Cluster-Robust Inference,” *Journal of Human Resources*, 50, 317–372.

Caplin, A. and D. Martin (2021): “Comparison of Decisions Under Unknown Experiments,” *Journal of Political Economy*, Forthcoming.

Carter, A. V., K. T. Schnepel, and D. G. Steigerwald (2017): “Asymptotic Behavior of at Test Robust to Cluster Heterogeneity,” *Review of Economics and Statistics*.

Charness, G. and D. Levin (2005): “When Optimal Choices Feel Wrong: A Laboratory Study of Bayesian Updating, Complexity, and Affect,” *American Economic Review*, 95, 1300–1309.

Chen, Y. (2011): “Perturbed communication games with honest senders and naive receivers,” *Journal of Economic Theory*, 146, 401–424.

Cooper, D. J. and J. H. Kagel (2016): “Other-Regarding Preferences: A Selective Survey of Experimental Results,” in *The Handbook of Experimental Economics, Volume 2*, ed. by J. H. Kagel and A. E. Roth, Princeton University Press.

Crawford, V. P., M. A. Costa-Gomes, and N. Iriberri (2013): “Structural Models of Nonequilibrium Strategic Thinking: Theory, Evidence, and Applications,” *Journal of Economic Literature*, 51:1, 5–62.

Crawford, V. P. and J. Sobel (1982): “Strategic Information Transmission,” *Econometrica*, 50, 1431–1451.

de Clippel, G. and K. Rozen (2020): “Communication, Perception, and Strategic Obfuscation,” *Working Paper*.

de Clippel, G. and X. Zhang (2020): “Non-Bayesian Persuasion,” *Working Paper*.

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

Dickhaut, J., K. McCabe, and A. Mukherji (1995): “An Experimental Study of Strategic Information Transmission,” *Economic Theory*, 6, 389–403.

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

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

Embrey, M., G. R. Fréchette, and S. Yuksel (2017): “Cooperation in the Finitely Repeated Prisoner’s Dilemma,” *Quarterly Journal of Economics*, 133, 509–551.

Forsythe, R., R. M. Isaac, and T. R. Palfrey (1989): “Theories and Tests of ”Blind Bidding” in Sealed-bid Auctions,” *RAND Journal of Economics*, 20, 214–238.

Forsythe, R., R. Lundholm, and T. Rietz (1999): “Cheap Talk, Fraud, and Adverse Selection in Financial Markets: Some Experimental Evidence,” *The Review of Financial Studies*, 12, 481–518.

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

Galor, E. (1985): “Information Sharing in Oligopoly,” *Econometrica*, 53, 329–343.

Gilligan, T. W. and K. Krehbiel (1987): “Decisionmaking and Standing Committees: An Informational Rationale for Restrictive Amendment Procedures,” *Journal of Law, Economics*, 3, 287–335.

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

Gneezy, U. (2005): “Deception: the role of consequences,” *American Economic Review*, 95(1), 384–394.

Goeree, J. K., C. A. Holt, and T. R. Palfrey (2016): *Quantal Response Equilibrium: A Stochastic Theory of Games*, Princeton University Press.

Green, J. R. and N. L. Stokey (2007): “A Two-person Game Of Information Transmission,” *Journal Of Economic Theory*, 135, 90–104.

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

Hagenbach, J., F. Koessler, and E. Perez-Richet (2014): "Certifiable Pre-Play Communication: Full Disclosure," *Econometrica*, 82(3), 1093–1131.

Hagenback, J. and E. Perez-Richet (2018): "Communication with Evidence in the Lab," *Working Paper*.

Hart, S., I. Kremer, and M. Perry (2017): "Evidence Games: Truth and Commitment," *American Economic Review*, 107(3), 690–713.

Hastie, T., R. Tibshirani, and J. Friedman (2009): *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer series in statistics New York, NY, USA:, second Edition.

Holt, C. A. (2007): *Markets, Games, and Strategic Behavior*, Pearson Addison Wesley Boston, MA.

Huck, S., W. Muller, and H.-T. Normann (2001): "Stackelberg Beats Cournot — On Collusion and Efficiency in Experimental Markets," *Economic Journal*, 111, 749–765.

Huck, S. and W. Müller (2000): "Perfect versus Imperfect Observability—An Experimental Test of Bagwell's Result," *Games and Economic Behavior*, 31, 174 – 190.

Hurkens, S. and N. Kartik (2009): "Would I lie to you? On social preferences and lying aversion," *Experimental Economics*, 12, 180–192.

Ibragimov, R. and U. K. Müller (2010): "t-Statistic Based Correlation and Heterogeneity Robust Inference," *Journal of Business & Economic Statistics*, 28, 453–468.

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

Jin, G., M. Luca, and D. Martin (2019): "Complex Disclosure," *Working Paper*.

——— (2020): "Is No News (Perceived As) Bad News? An Experimental Investigation of Information Disclosure," *American Economic Journal: Microeconomics*.

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

Kamenica, E. (2019): "Bayesian Persuasion and Information Design," *Annual Review of Economics*, 11, 249–272.

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

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

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

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

Lipnowski, E., D. Ravid, and D. Shishkin (2018): "Persuasion via Weak Institutions," *Working Paper*.

MacQueen, J. (1967): "Some Methods for Classification and Analysis of Multivariate Observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Oakland, CA, USA, vol. 1, 281–297.

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

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

——— (2008): "What the Seller Won't Tell You: Persuasion and Disclosure in Markets," *Journal of Economic Perspectives*, 22, 115–131.

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

Morgan, J. and F. Várdy (2004): "An experimental study of commitment in Stackelberg games with observation costs," *Games and Economic Behavior*, 49, 401 – 423.

——— (2013): "The Fragility of Commitment," *Management Science*, Vol. 59, No. 6, 1344–1353.

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., A. Postlewaite, and K. Suzumura (1990): "Strategic Information Revelation," *The Review of Economic Studies*, 57, 25–47.

Perez-Richet, E. and V. Skreta (2018): "Test Design under Falsification," *Working Paper*.

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

Sánchez-Pagés, S. and M. Vorsatz (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., M. Spezio, and C. Camerer (2010): "Pinocchio's Pupil: Using Eyetracking and Pupil Dilation to Understand Truth Telling and Deception in Sender-Receiver Games," *American Economic Review*, 100, 984–1007.

Wilson, A. and E. Vespa (2020): "Information Transmission Under the Shadow of the Future: An Experiment," *American Economic Journal: Microeconomics*, Forthcoming.

A Equilibrium Refinement and Proofs

In this section, we formally present the refinement, illustrate two examples of PBE that fail it, and characterize the set of equilibria that survive it. We begin with a formal definition of PBE for our framework. Recall that, in the commitment stage, the sender chooses an information structure $\pi_C \in \Pi$. Then, at every history π'_C, the sender observes θ and chooses a revision strategy, denoted $\zeta_R(\pi'_C) \in \Pi$, which possibly depends on π'_C. In the last stage, the receiver observes the history (m, π'_C) and responds by guessing $a \in \{a_H, a_L\}$. We denote her (possibly mixed) strategy by $\sigma(m, \pi'_C)$. A system of beliefs μ assigns a posterior probability to θ_H conditional on every message m, possibly as a function of π'_C and $\zeta_R(\pi'_C)$.

Definition 1. *Fix* (Π, ρ, q). *The tuple* $(\pi_C, \zeta_R, \sigma, \mu)$ *is a Perfect Bayesian Equilibrium if*

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

(2) For all (π'_C, θ), $\sum_m \zeta_R(\pi'_C)(m|\theta)v(\sigma(m, \pi'_C)) \geq \sum_m \pi_R(m|\theta)v(\sigma(m, \pi'_C))$ *for all* π_R*;*

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

(4) For all (m, π'_C), *the posterior belief* $\mu(m, \pi'_C, \zeta_R(\pi'_C))$ *is computed given* $\rho\pi'_C + (1-\rho)\zeta_R(\pi'_C)$ *using Bayes' rule whenever possible.*[39]

We refine the set of PBE by assuming that, in both the commitment and the revision stage, the sender breaks indifference in favor of strategies that send message $m = \theta_H$ conditional on θ_H with higher probability. More formally, a PBE $(\pi^*_C, \zeta^*_R, \sigma^*, \mu^*)$ satisfies the refinement if the following holds. In the revision stage, at any history (π'_C, θ_H), if there is a strategy π'_R that leads to the same continuation payoff as $\zeta^*_R(\pi'_C)$, then $\zeta^*_R(\pi'_C)(\theta_H|\theta_H) \geq \pi'_R(\theta_H|\theta_H)$. For example, if there is a message $m \neq \theta_H$ such that $\sigma^*(m, \pi'_C) = \sigma^*(\theta_H, \pi'_C)$, then $\zeta^*_R(\pi'_C)(m|\theta_H) = 0$. In the commitment stage, if there is a strategy π'_C that leads to the same continuation payoff as π^*_C, then $\rho\pi^*_C(\theta_H|\theta_H) + (1-\rho)\zeta^*_R(\pi^*_C)(\theta_H|\theta_H) \geq \rho\pi'_C(\theta_H|\theta_H) + (1-\rho)\zeta^*_R(\pi'_C)(\theta_H|\theta_H)$.

The idea behind our refinement rests on two forces. On the one hand, the sender may suffer a small psychological cost when not telling the truth. Thus, whenever indifferent, she could break ties in favor of being honest. On the other hand, the sender may believe that a small fraction of receivers is naive and reads messages at face value. That is, these receivers respond to message θ_H by guessing a_H. Thus, whenever indifferent, the sender may break ties in favor

[39] Recall that, when information is verifiable, we assume that $\mu(\theta_H, \pi_C, \pi_R) = 1$ and $\mu(\theta_L, \pi_C, \pi_R) = 0$, for all π_C and π_R. We find this assumption in the spirit of Battigalli and Siniscalchi (2002).

of sending message θ_H regardless of the state. Similar forces have been considered in the literature (e.g., see Chen (2011) and Hart et al. (2017); for experiments, see Cai and Wang (2006) and Blume et al. (2020)) and are especially prominent in experimental settings like ours, in which messages are coded with literal meanings. It is difficult from an abstract perspective to evaluate the weight that a sender may give to each of these two forces. However, conditional on state θ_H, the two forces go in the same direction, and thus their effect is unambiguous: By sending message θ_H, the θ_H-type sender is both honest and opportunistic. We view this as a justification for assuming that, in this case, the sender will break ties in favor of sending such a message. Instead, conditional on the state θ_L, the sender could break ties either by being honest (i.e., $m = \theta_L$) or by sending the opportunistic message (i.e., $m = \theta_H$). The effect is ambiguous. In this case, it seems reasonable to impose no restriction and let the sender randomize if so she desires. In a nutshell, we imagine our senders thinking that it cannot hurt to tell the truth when it is convenient.[40]

In the rest of this appendix, we refer to this tie-breaking rule with the acronym TWC, which stands for "truthful when convenient." A TWC equilibrium is a PBE that satisfies TWC. The next result shows that this tie-breaking rule is powerful enough to select a unique equilibrium outcome for each ρ. Throughout this appendix, we will make repeated use of the two thresholds introduced in Section 2.2, namely $\underline{\rho} := \frac{q-\mu_0}{q(1-\mu_0)}$ and $\bar{\rho} = \frac{q(1-\mu_0)}{q(1-\mu_0)+(1-q)\mu_0}$. Moreover, we say that an equilibrium achieves *full-commitment correlation* (FCC) if the state-action correlation is equal to $\sqrt{q\underline{\rho}}$. This benchmark is the correlation achieved by any PBE under full commitment and unverifiable information.

Theorem 1. *TWC equilibria exist.*

> *(Unverifiable) If $\rho < \underline{\rho}$, all TWC equilibria have zero correlation. If $\rho \geq \underline{\rho}$, they all achieve FCC.*
>
> *(Verifiable) If $\rho < \bar{\rho}$, all TWC equilibria have correlation one. If $\rho \geq \bar{\rho}$, their correlation is equal to $(\frac{q(1-\rho(1-\underline{\rho})}{q+\rho(1-q)})^{\frac{1}{2}}$.*

The proof of this result is in Appendix A.1.

Finally, we present two examples—for unverifiable and verifiable information, respectively—that indicate why Theorem 1 can fail without the tie-breaking rule imposed by our refinement.

[40]Our data provide support for this refinement. First, we find that message r has a significant, albeit small, positive effect on the probability that the receiver guesses *red*, even when controlling for the induced posterior. Second, in the revision stage, senders send message r conditional on R with a median probability equal to 1.

These examples illustrate that, in the absence of a refinement, there are equilibria that feature behavior that is somewhat unreasonable.

Example 1: *Unverifiable Information.*

Let information be unverifiable. Assume $\rho = \frac{3}{5}$, $q = \frac{1}{2}$, and $\mu_0 = \frac{1}{3}$. Note that, in this case, $\rho > \underline{\rho}$. Consider the pair of sender's strategies (π_C, π_R) in Table A5. Given these strategies, note that beliefs satisfy $\mu(\theta_H, \pi_C, \pi_R) < q$ and $\mu(\theta_L, \pi_C, \pi_R) < q$. That is, despite π_C being fully revealing, the sender's behavior in the revision stage entirely garbles the information transmitted in the commitment stage.

Table A5

π_C	$m = \theta_H$	$m = \theta_L$	$m = n$
θ_H	1	0	0
θ_L	0	1	0

π_R	$m = \theta_H$	$m = \theta_L$	$m = n$
θ_H	0	1	0
θ_L	1	0	0

When $\rho = \frac{3}{5}$, it can be shown that for all commitment strategies π'_C, there exists a retaliatory strategy π'_R, similar to the one from Table A5, that garbles the information contained in π'_C. That is, the pair (π'_C, π'_R) induces the receiver to choose a_L conditional on all messages. This means that a PBE with correlation zero exists, even if, in this case, $\rho > \underline{\rho}$. Similarly, we can show that a PBE with correlation higher than FCC exists. The particularly strange behavior that characterizes these PBE is ruled out by the TWC refinement. For example, consider the history in which the pair of strategies in Table A5 is played by sender. As argued, the θ_H-type sender in the revision stage is indifferent between sending message θ_H and θ_L, given that both lead to action a_L. In this case, the refinement requires that the sender breaks ties in favor of message θ_H, that is, sets $\pi_R(\theta_H|\theta_H) = 1$.

Example 2: *Verifiable Information.*

Now assume that information is verifiable. As above, let $\rho = \frac{3}{5}$, $q = \frac{1}{2}$, and $\mu_0 = \frac{1}{3}$. Consider the pair of strategies (π_C, π_R) that is described in Table A6. Conditional on π_C, there exists a continuation PBE in which π_R is played, and $\sigma(m) = a_H$ if $m \in \{\theta_H, n\}$ and a_L otherwise. In such a continuation equilibrium, the sender of type θ_H is indifferent between the two feasible messages θ_H and n, as they both lead to a_H (see footnote 39). Note that the profile of strategies (π_C, π_R, σ) achieves FCC. This PBE, however, fails the TWC refinement. Indeed, the θ_H-type sender is indifferent in the revision stage between sending message n and the verifiable message θ_H. In this case, the refinement requires that the sender breaks ties in favor of message θ_H, that is, sets $\pi_R(\theta_H|\theta_H) = 1 \neq 0$

Table A6

π_C	$m=\theta_H$	$m=\theta_L$	$m=n$		π_R	$m=\theta_H$	$m=\theta_L$	$m=n$
θ_H	0	0	1		θ_H	0	0	1
θ_L	0	$\frac{5}{6}$	$\frac{1}{6}$		θ_L	0	0	1

A.1 Proofs

Proof of Theorem 1.

Unverifiable Information.

Case $\rho < \underline{\rho}$. Let information be unverifiable and $\rho < \underline{\rho}$. We begin by showing that no TWC equilibrium can have nonzero correlation. Let $(\pi_C, \zeta_R, \sigma, \mu)$ be a TWC equilibrium. Let $\pi_R = \zeta_R(\pi_C)$ and $\sigma = (\cdot, \pi_C)$. Suppose by way of contradiction that $\phi(\pi_C, \pi_R, \sigma) > 0$. Since $q > \mu_0$, this implies that action a_H is chosen with strictly positive probability. Let $\varnothing \neq \bar{M} \subsetneq M$ be the set of messages such that $\sigma(m, \pi_C) = a_H$, for $m \in \bar{M}$. Note that condition (2) in Definition 1 implies that $\sum_{m\in\bar{M}} \pi_R(m|\theta) = 1$ for all θ. Similarly, condition (3) implies that $\mu(m, \pi_C, \pi_R) \geq q$, for $m \in \bar{M}$. Note that the probability of θ_H conditional on receiving a message in $\bar{M}$ is

$$\begin{aligned}
\Pr(\theta_H|\bar{M}) &= \frac{\mu_0(\rho\sum_{m\in\bar{M}}\pi_C(m|\theta_H) + (1-\rho))}{\mu_0(\rho\sum_{m\in\bar{M}}\pi_C(m|\theta_H) + (1-\rho)) + (1-\mu_0)(\rho\sum_{m\in\bar{M}}\pi_C(m|\theta_L) + (1-\rho))} \\
&\leq \frac{\mu_0}{\mu_0 + (1-\mu_0)(1-\rho)} \\
&< \frac{\mu_0}{\mu_0 + (1-\mu_0)(1-\underline{\rho})} = q.
\end{aligned}$$

The first equality follows from Bayes' rule. The first inequality holds because $\Pr(\theta_H|\bar{M})$ is maximized when $\sum_{m\in\bar{M}} \pi_C(m|\theta_H) = 1$ and $\sum_{m\in\bar{M}} \pi_C(m|\theta_L) = 0$. The third inequality holds because $\rho < \underline{\rho}$. The last equality can be verified by substituting the expression for $\underline{\rho}$. However, Bayes' rule implies that, for appropriately chosen positive weights $(\beta_m)_{m\in\bar{M}}$, $\Pr(\theta_H|\bar{M}) = \sum_{m\in\bar{M}} \beta_m \mu(m, \pi_C, \pi_R)$.[41] Since, by assumption, $\mu(m, \pi_C, \pi_R) \geq q$ for $m \in \bar{M}$, we have $\Pr(\theta_H|\bar{M}) \geq q$, a contradiction. Therefore, $\bar{M} = \varnothing$, that is $\sigma(m, \pi_C) = a_L$ for all m. This implies that $\phi(\pi_C, \pi_R, \sigma) = 0$.

We are left to show that a TWC equilibrium $(\pi_C, \zeta_R, \sigma, \mu)$ exists. Fix any history π'_C. When $\rho < \underline{\rho}$, Lemma 2 in Appendix D.1 can be specialized to show that the continuation TWC equilibrium $(\pi_R, \sigma(\cdot, \pi'_C), \mu(\cdot, \pi'_C, \pi_R))$ given π'_C that this lemma constructs is such that $\sigma(m, \pi'_C) = a_L$

[41] More specifically, if $\bar{M} = \{m'\}$, then $\beta_{m'} = 1$; if $\bar{M} = \{m', m''\}$, then $\beta_{m'} := \frac{\sum_\theta \mu_0(\theta)(\rho\pi_C(m'|\theta)+(1-\rho)\pi_R(m'|\theta))}{\sum_\theta \mu_0(\theta)(\rho\sum_{m\in\bar{M}}\pi_C(m|\theta)+(1-\rho))}$ and $\beta_{m''} = 1 - \beta_{m'}$.

for all m.[42] This implies that for all π'_C, the sender earns 0. Therefore, she is indifferent among all π'_C. By the TWC refinement, the sender breaks indifference as follows. Let $(x, y) \in [0, 1]$ satisfy $(1 - \underline{\rho}) < \rho x + (1 - \rho)y$. The sender chooses π_C defined as $\pi_C(\theta_H|\theta_H) = 1$, $\pi_C(\theta_H|\theta_L) = x$, and $\pi_C(\theta_L|\theta_L) = x'$. This leads to a revision stage in which she chooses $\zeta_R(\pi_C) = \pi_R$ defined as $\pi_R(\theta_H|\theta_H) = 1$, $\pi_R(\theta_H|\theta_L) = y$, and $\pi_R(\theta_L|\theta_L) = y'$.

Case $\rho \geq \underline{\rho}$. We first show that all TWC equilibria must achieve FCC. Suppose by contradiction that there is a TWC equilibrium $(\pi_C, \zeta_R, \sigma, \mu)$ such that, letting $\pi_R = \zeta_R(\pi_C)$ and $\sigma = \sigma(\cdot, \pi_C)$ be the on-path strategies, $\phi(\pi_C, \pi_R, \sigma) \neq \sqrt{q\underline{\rho}}$. We want to show that the sender has a profitable deviation $\hat{\pi}$ in the commitment stage. Let $\hat{\pi}_C(\theta_H|\theta_H) = 1$, $\hat{\pi}_C(\theta_H|\theta_L) = x$, and $\hat{\pi}_C(\theta_L|\theta_L) = 1 - x$, where $x := \frac{1}{\rho}(\rho - \underline{\rho}) \in [0, 1]$. Let $\hat{\pi}_R = \zeta_R(\hat{\pi}_C)$. Let $\hat{M} := \{m \mid \mu(m, \hat{\pi}_C, \hat{\pi}_R) \geq q$ and m has positive probability$\}$. First, we show that $\hat{M} \neq \emptyset$. Suppose that is not the case. If $\hat{M} = \emptyset$, $\sigma(m, \hat{\pi}_C) = a_L$ for all m and, thus, the θ_H-type sender in the revision stage is indifferent between all messages. Since the equilibrium satisfies TWC, $\hat{\pi}_R(\theta_H|\theta_H) = 1$. However,

$$\mu(\theta_H, \hat{\pi}_C, \hat{\pi}_R) = \frac{\mu}{\mu + (1 - \mu)(\rho x + (1 - \rho)\pi_R(\theta_H|\theta_L))} \geq \frac{\mu}{\mu + (1 - \mu)(\rho x + (1 - \rho))} = q.$$

Therefore, $\hat{M} \neq \emptyset$. Next, we show that $\theta_H \in \hat{M}$. By way of contradiction, suppose instead that $\theta_H \notin \hat{M}$. Then, since $\hat{M} \neq \emptyset$, it must be that $\hat{\pi}_R(\theta_H|\theta) = 0$ for all θ (Condition 2 in Definition 1). However, in this case,

$$\mu(\theta_H, \hat{\pi}_C, \hat{\pi}_R) = \frac{\rho\mu}{\rho\mu + (1 - \mu)\rho x} \geq q.$$

Therefore, $\theta_H \in \hat{M}$. Finally, we show that $\hat{M} = \{\theta_H\}$. Note that, since $q > \mu_0$, $\hat{M} \subsetneq M$. Therefore, suppose $m' \neq \theta_H$ and $m' \in \hat{M}$. Since the equilibrium is truth-leaning, $\hat{\pi}_R(\theta_H|\theta_H) = 1$. Let $\hat{\pi}_R(m'|\theta_L) = a'$ and $\hat{\pi}_R(\theta_H|\theta_L) = 1 - a'$. If $m' = \theta_L$, $\mu(\theta_L, \hat{\pi}_C, \hat{\pi}_R) = 0$ and, thus, $m' \notin \hat{M}$. If $m' = n$, then either m' has zero probability (if $a' = 0$) or $m' = \theta_L$, $\mu(\theta_L, \hat{\pi}_C, \hat{\pi}_R) = 0$ (if $a' > 0$). In either case, $m' \notin \hat{M}$. Therefore, we conclude that $\hat{M} = \{\theta_H\}$. This uniquely pins down the revision strategy $\hat{\pi}_R$, which is $\hat{\pi}_R(\theta_H|\theta) = 1$, for all θ. Letting $\hat{\sigma} = \sigma(\cdot, \hat{\pi}_C)$ is easy to verify that $\phi(\hat{\pi}_C, \hat{\pi}_R, \hat{\sigma}) = \sqrt{q\underline{\rho}}$ and that $\hat{\pi}_C$ leads to a continuation equilibrium in which the sender earns her first-best payoff μ_0/q. In contrast, the sender expects to earn a strictly lower payoff on the equilibrium path of $(\pi_C, \zeta_R, \sigma, \mu)$. This is because, by assumption, $\phi(\pi_C, \pi_R, \sigma) \neq \sqrt{q\underline{\rho}}$, which implies (see Lemma 1 in Appendix D.1) that the sender earns a payoff strictly lower than μ_0/q.

[42] In reference to the three cases discussed in 2, note that: Case 1 and Case 3.ii lead to $\sigma(m, \pi'_C) = a_L$ for all m with no further qualification; For Case 2.(i) note that, if $\rho < \underline{\rho}$, $\pi_R(\theta_H|\theta) = 1$ for all θ implies that $\mu(\theta_H, \pi_C, \pi_R) < q$, regardless of π'_C. Therefore, in this case $\sigma(m, \pi'_C) = a_L$ for all m; Finally, in Case 2.(ii) and Case 3.(i), $\rho < \underline{\rho}$ implies that $\Lambda' < (1 - \rho)\underline{\rho}/\rho$ and $\delta^* < \delta_*$. Therefore, $\sigma(m, \pi'_C) = a_L$ for all m.

Therefore, $\hat{\pi}_C$ is a strictly profitable deviation and, thus, $(\pi_C, \zeta_R, \sigma, \mu)$ is not a TWC equilibrium.

Next, we show that a TWC equilibria exists. For each history π'_C, let us define a continuation TWC equilibrium as described in Lemma 2. On the equilibrium path, instead, the sender chooses $\hat{\pi}_C$, as was defined above. This strategy leads to $\zeta_R(\hat{\pi}_C) = \hat{\pi}_R$, again as defined above. Note that these two strategies have $\pi_C(\theta_H|\theta_H) = \pi_R(\theta_H|\theta_H) = 1$ and, thus, they (trivially) satisfy the TWC refinement. Conditional on these strategies, the receiver chooses a_H if $m = \theta_H$ and chooses a_L otherwise. Finally, the receiver's beliefs are pinned down by Bayes' rule if $m \in \{\theta_H, \theta_L\}$ and are equal to zero otherwise. It is easy to verify that these strategies define a TWC equilibrium.

Finally, suppose that $(\pi'_C, \zeta'_R, \sigma', \mu')$ is another TWC equilibrium. We argue that $\pi'_C(\theta_H|\theta_H) = 1$. Suppose that is not the case. Above, we showed that all TWC equilibria achieve FCC. This implies that, in the commitment stage, the sender is indifferent between playing π'_C or deviating to $\hat{\pi}_C$. Since $\pi'_C(\theta_H|\theta_H) < 1$ she breaks ties in favor of $\hat{\pi}_C$, which instead has $\hat{\pi}_R(\theta_H|\theta) = 1$. Therefore, $(\pi'_C, \zeta'_R, \sigma', \mu')$ does not satisfy TWC, a contradiction. □

Verifiable Information.

We begin by proving an ancillary result. Fix $\rho \in [0, 1)$. For every π_C, we want to show that there exists a continuation TWC equilibrium (π_R, σ, μ). We do so by construction. Fix any π_C. Verifiability requires that $\mu(\theta_H, \pi_C, \pi_R) = 1$ and $\mu(\theta_L, \pi_C, \pi_R) = 0$ (see footnote 39). Therefore, $\sigma(\theta_H, \pi_C) = a_H$ and $\sigma(\theta_L, \pi_C) = a_L$ and therefore, the TWC refinement requires that $\pi_R(\theta_H|\theta_H) = 1$. We are left to determine $\delta := \pi_R(n|\theta_L)$, $\sigma(n)$, and $\mu(n, \pi_C, \pi_R)$. To simplify notation, let $\pi_C(n|\theta_H) = x$, $\pi_C(n|\theta_L) = y$. Note that, since information is verifiable, π_C is uniquely pinned down by $(x, y) \in [0, 1]^2$. Define $\Phi = \frac{\rho}{1-\rho}((1-\underline{\rho})x - y)$. If $\Phi \geq 1$, we let $\delta = 1$ and $\sigma(n, \pi_C) = a_H$. In this case, it is easy to verify that $\mu(n, \pi_C, \pi_R) \geq q$, which is pinned down by Bayes' rule. If $\Phi \in [0, 1)$, we let $\delta \in (\Phi, 1]$ and $\sigma(n, \pi_C) = a_L$. It is easy to verify that, in this case, $\mu(n, \pi_C, \pi_R) < q$, which is pinned down by Bayes' rule. If $\Phi < 0$, we let $\delta \in [0, 1]$ and $\sigma(n, \pi_C) = a_L$. Then, once again, $\mu(n, \pi_C, \pi_R) < q$, which is pinned down by Bayes' rule. Finally, there is one more continuation equilibrium to discuss. If $x = y = 0$ (and, thus, $\Phi = 0$), we let $\delta = 0$ and $\sigma(n, \pi_C) = a_L$. In this case, we can set $\mu(n, \pi_C, \pi_R) < q$, which is not pinned down by Bayes' rule. In each of the cases above, it is straightforward to verify that the triple (π_R, σ, μ) is a continuation TWC equilibrium given π_C.

We now prove the statement of the Theorem.

Case $\rho < \bar{\rho}$. We begin by showing that a TWC equilibrium $(\pi^*_C, \zeta^*_R, \sigma^*, \mu^*)$ exists. We do so by construction. For all π_C, note that $\Phi < 1$. Indeed, $\Phi = \frac{\rho}{1-\rho}((1-\underline{\rho})x - y) \leq \frac{\rho}{1-\rho}(1-\underline{\rho})$, since Φ is maximized when $x = 1$ and $y = 0$, and $\frac{\rho}{1-\rho}(1-\underline{\rho}) < 1$ if $\rho < \frac{1}{2-\underline{\rho}} = \bar{\rho}$. Therefore, the argument above allows us to pin down a continuation TWC equilibrium (π_R, σ, μ) given every π_C. Since $\Phi < 1$, for each π_C $\sigma(m) = a_H$ if and only if $m = \theta_H$. Therefore $V = \sum_{\theta,m} \mu_0(\theta)(\rho\pi_C(m|\theta) + (1-\rho)\pi_R(m|\theta))v(\sigma(m)) = \mu_0(1-\rho x)$. In the commitment stage, it is thus optimal to set $x^* = 0$ and $y^* \in [0,1]$. This characterizes the enitre set of TWC equilibria. Moreover, it is straightforward to verify that all these equilibria have correlation 1, since the receiver plays a_H if and only if $\theta = \theta_H$.

Case $\rho \geq \bar{\rho}$. Suppose π_C is such that $\Phi < 1$. Then, as before, the sender's expected payoff is $V^{<1} = \mu_0(1-\rho x)$. Suppose, instead, π_C is such that $\Phi \geq 1$. Then, since in this case $\delta = 1$ and $\sigma(n) = a_H$, $V^{\geq 1} = \sum_{\theta,m} \mu_0(\theta)(\rho\pi_C(m|\theta) + (1-\rho)\pi_R(m|\theta))v(\sigma(m)) = \mu_0 + (1-\mu_0)(1-\rho(1-y))$. Note that $V^{\geq 1} > V^{<1}$. Therefore, equilibrium behavior in the commitment stage requires that the sender maximizes y while satisfying $\Phi \geq 1$. Since Φ is decreasing in y, this entails setting $\Phi = 1$, which leads to setting $x = 1$ and $y = 1 - \underline{\rho} - \frac{1-\rho}{\rho}$. Note that $y \geq 0$ since $\rho \geq \bar{\rho}$. Summing up, when $\rho \geq \bar{\rho}$, on the equilibrium path of all TWC equilibria, the sender plays the same strategies (π_C^*, π_R^*), characterized by $x^* = 1$, $y^* = 1 - \underline{\rho} - \frac{1-\rho}{\rho}$, and $\delta^* = 1$, while the receiver responds with $\sigma^*(m, \pi_C^*) = a_L$ if $m = \theta_L$ and a_H otherwise. Therefore, all such equilibria induce the same correlation, which is $\phi(\pi_C^*, \pi_R^*, \sigma^*) = (\frac{q-\mu_0(\rho+q(1-\rho))}{(1-\mu_0)(\rho+q(1-\rho))})^{\frac{1}{2}}$, or equivalently $(\frac{q(1-\rho(1-\underline{\rho}))}{q+\rho(1-q)})^{\frac{1}{2}}$. □

Proof of Proposition 1. Suppose that information is unverifiable and $1 > \rho \geq \underline{\rho}$. Let $(\pi_C, \zeta_R, \sigma, \mu)$ be a TWC equilibrium. The proof of Theorem 1 establishes that $\pi_C(\theta_H|\theta_H) = 1$. Let $\bar{M}$ be the set of messages inducing action a_H in equilibrium. Since the equilibrium correlation is strictly positive (Theorem 1), $\bar{M} \neq \emptyset$, that is, at least one message leads to action a_H. Moreover, since $\pi_C(\theta_H|\theta_H) = 1$, $\theta_H \in \bar{M}$. Thus, the TWC refinement requires that the on-path revision strategy π_R satisfies $\pi_R(\theta_H|\theta_H) = 1$. This implies that $\bar{M} = \{\theta_H\}$ and, thus, $\pi_R(\theta_H|\theta_L) = 1$. In turn, this implies that, in the commitment stage, the sender finds it optimal to choose $\pi_C(\theta_H|\theta_L) = \frac{\rho-\underline{\rho}}{\rho}$, leading to $\mu(\theta_H, \pi_C, \pi_R) = q$. Letting $\delta \in [0,1]$, the remainder of the commitment strategy is given by $\pi_C(\theta_L|\theta_L) = \frac{\delta\underline{\rho}}{\rho}$ and $\pi_C(\theta_L|\theta_L) = \frac{(1-\delta)\underline{\rho}}{\rho}$. Using the definition of ϕ^B, it is straightforward to compute the Bayesian correlations induced by π_C and π_R respectively. Consider the commitment stage strategy π_C. For all δ, a hypothetical Bayesian receiver would choose a_H conditional on $m = \theta_H$ and a_L otherwise. Therefore, $\phi^B(\pi_C) = (\frac{\mu_0\underline{\rho}}{\rho-\underline{\rho}(1-\mu_0)})^{\frac{1}{2}}$. Note that $\phi^B(\pi_C) > 0$ since $\rho \geq \underline{\rho}$. Consider the revision stage. Since $\pi_R(\theta_H|\theta) = 1$ for all θ, a hypothetical Bayesian receiver would choose a_L conditional on all messages. Thus, $\phi^B(\pi_R) = 0$. We conclude that $\phi^B(\pi_C) > \phi^B(\pi_R)$.

Suppose instead that information is verifiable and let $1 > \rho > \bar{\rho}$. Let $(\pi_C, \zeta_R, \sigma, \mu)$ be a TWC equilibrium. The proof of Theorem 1 shows that, on the equilibrium path of any TWC equilibrium the sender plays the same strategies (π_C, π_R). For the commitment stage, we have that $\pi_C(n|\theta_H) = 1$, $\pi_C(n|\theta_L) = 1 - \underline{\rho} - \frac{1-\rho}{\rho}$. Given π_C, a hypothetical Bayesian receiver would guess a_H conditional on receiving message θ_H or n, and she would guess a_L otherwise. Therefore, it is easy to verify that $\phi^B(\pi_C) = \frac{\mu_0(1-\rho(1-\underline{\rho}))}{\rho-(1-\mu_0)(1-\rho(1-\underline{\rho})}$. Note that $\phi^B(\pi_C) < 1$ since $\rho > \bar{\rho}$. For the commitment stage, we have $\pi_R(\theta_H|\theta_H) = 1$ and $\pi_R(n|\theta_L) = 1$. Given such a π_R, a hypothetical Bayesian receiver would guess a_H conditional on receiving message θ_H and a_L otherwise. It is immediate to see that $\phi^B(\pi_R) = 1$. We conclude that $\phi^B(\pi_C) < \phi^B(\pi_R)$. □

Proof of Proposition 2. The statement of the proposition directly follows from Theorem 1. When information is unverifiable, we established that, letting $(\pi_C, \zeta_R, \mu, \sigma)$ be a TWC equilibrium and (π_C, π_R, σ) the strategy profile that is played on the equilibrium path,

$$\phi(\pi_C, \pi_R, \sigma) = \begin{cases} 0 & \text{if } \rho < \underline{\rho} \\ \sqrt{q\underline{\rho}} & \text{if } \rho \geq \underline{\rho}. \end{cases}$$

Therefore, when information is unverifiable, the equilibrium correlation weakly increases in ρ. Conversely, assume that information is verifiable. Theorem 1 established that, letting $(\pi_C, \zeta_R, \mu, \sigma)$ be a TWC equilibrium and (π_C, π_R, σ) the strategy profile that is played on the equilibrium path,

$$\phi(\pi_C, \pi_R, \sigma) = \begin{cases} 1 & \text{if } \rho < \bar{\rho} \\ (\frac{q(1-\rho(1-\underline{\rho})}{q+\rho(1-q)})^{\frac{1}{2}} & \text{if } \rho \geq \bar{\rho}. \end{cases}$$

It is easy to verify that $\frac{q(1-\rho(1-\underline{\rho})}{q+\rho(1-q)}$ is decreasing in ρ. Finally, consider the extreme case where $\rho = 1$. In this case, $(\frac{q(1-\rho(1-\underline{\rho})}{q+\rho(1-q)})^{\frac{1}{2}} = \sqrt{q\underline{\rho}}$. □

Online Appendix for

Rules and Commitment in Communication: An Experimental Analysis

Guillaume Fréchette
New York University

Alessandro Lizzeri
New York University

Jacopo Perego
Columbia University

Contents

B Additional Treatments

B.1 $U100H$ – Changing Receiver's Incentives

In this section, we test a different comparative static result: instead of varying the degree of commitment or the communication rules, we change the alignment between the sender's and the receiver's preferences. More precisely, we increase the persuasion threshold q. As we explain below, this can be done experimentally by changing the preferences of the receiver. Formally, the prediction that we test is the following.

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

This result shows that when ρ is sufficiently high, an increase in q increases equilibrium correlation, irrespective of the communication rules. In particular, when $\rho = 1$, raising q strictly increases the equilibrium correlation for both verifiability scenarios.

Based on this idea, we designed an additional treatment with full commitment ($\rho = 1$) and unverifiable information. We label this treatment $U100H$ and compare it directly to $U100$.[2] Payoffs are as follows. As in all other treatments, the receiver earns nothing if she guesses incorrectly. In contrast to our main treatments however, the receiver earns \$2 if she correctly guesses that $\theta = B$, but only 67¢ if she correctly guesses that $\theta = R$. This payoff structure increases the persuasion threshold from $q = 1/2$ to $q = 3/4$. Since the receiver is harder to persuade, the sender is automatically worse off relative to $U100$. Therefore, to improve the comparability between treatments, we also modify the sender's payoff in $U100H$. In particular, she earns \$3 (instead of \$2) whenever $a = red$. In this way, her expected equilibrium payoff is the same for $U100$ and $U100S$. In equilibrium, the sender chooses $\pi_C(r|R) = 1$ and $\pi_C(b|B) = 5/6$ and the predicted Bayesian correlation is $\phi^B(\pi_C) = 5/\sqrt{40} \approx 0.79$.

The left panel of Figure B9 reports the main clusters of senders' behavior in treatment $U100H$. These are computed through a k-means algorithm, as described in Section 5.2. A large fraction of senders, indicated by a blue square, choose strategies that are close to equilibrium behavior. A smaller but significant fraction of senders, indicated by a purple star, choose a strategy that would be close to equilibrium behavior in $U100$ but is not informative enough to persuade a Bayesian receiver in $U100H$. The strategies summarized by the red circle capture commitment blindness, while those summarized by the green diamond capture a cluster of residual strategies that should be interpreted as noise. When comparing these clusters with those computed for treatment $U100$

[2] We conducted four sessions of $U100H$, each with 16–20 subjects (72 in total). The sessions lasted approximately 100 minutes. Subjects earned on average \$32, including a show-up fee of \$10. On average, senders and receivers made \$23 and \$40, respectively.

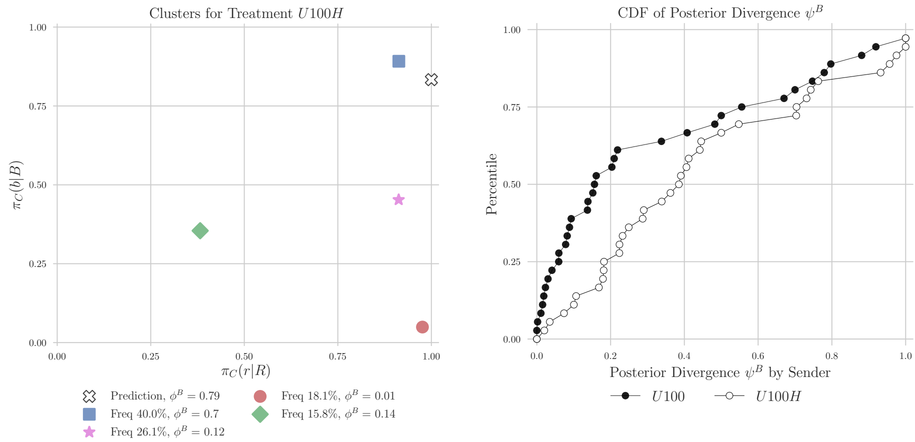


Figure B9: Strategy Clusters (left) and CDFs of Posterior Divergence ψ^B (right)

(Figure D18, right panel) or $U80$ (Figure 7), we observe an overall shift toward more-informative strategies, as predicted by the theory (upper-right corner).

Quantifying this shift is complicated by the fact that receivers' preferences between $U100$ and $U100H$ have changed. Therefore, Bayesian correlations ϕ^B have to be computed using different utilities for the receiver in the two treatments. For example, a posterior of 0.74 leads to $a = red$ for the Bayesian receiver of treatment $U100$, but $a = blue$ for that of treatment $U100H$. To avoid this problem, we measure information sent using ψ^B, the divergence between the expected posterior conditional on the states that we introduced in Section 4.2. Recall that ψ^B is proportional to the variance of the induced posteriors (see Online Appendix D.2). As such, it is independent of u and, thus, it is a more appropriate measure when comparing data from treatments that feature different q's. The divergence ψ^B in $U100$ is 0.30 (predicted 0.25); in $U100H$, it is 0.42 (predicted 0.63). The increase from $U100$ to $U100H$ is significant ($p < 0.01$), in line with Proposition 3. Moreover, the sender-by-sender CDF of ψ^B increases from $U100$ to $U100H$ in a first-order stochastic sense, as reported in the right panel of Figure B9.

Finally, the comparison between $U100$ and $U100H$ also speaks to the question of the relationship between subjects' behavior and the complexity of our design. Although the complexity of the senders' task changes between the commitment and the revision stages and perhaps even with varying levels of commitment and communication rules this complexity should be the same in $U100$ and $U100H$. Therefore, this comparison, in which the data corroborate the theoretical prediction of Proposition 3, should be immune to a "complexity critique."

B.2 $U100S$ – Simplifying the Message Space

In our main treatments, senders can choose among three messages: r, b, and n. In theory, when information is unverifiable, one of these messages is redundant and its presence does not change the equilibrium outcome. From a design perspective, message n is important as it allows a clean

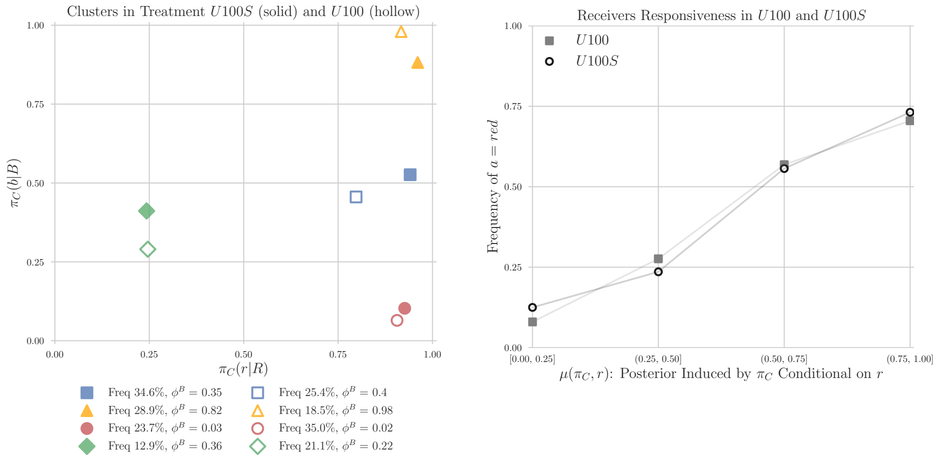


Figure B10: Senders' (left panel) and Receivers' (right panel) Behavior in $U100$ and $U100S$

comparison between treatments with and without verifiable information. In this section, we explore the effect of removing this redundant message in a treatment with unverifiable information and full commitment. Every other aspect of this treatment, which we label $U100S$, is identical to $U100$.[3] Implicitly, this is also a test of how the complexity of subjects' tasks affect their behavior. It is reasonable to think that treatment $U100$ is more complex than $U100S$ for both senders and receivers. If complexity was a major factor affecting subjects' behavior, one would expect to see differences in $U100S$ and $U100$. Our main conclusion from the comparison of $U100$ and $U100S$ is that adding message n increases the noise but does not significantly alter the average behavior.

We begin by comparing the senders' behavior in treatments $U100$ and $U100S$. The left panel of Figure B10 reports the main clusters for these treatments computed through a k-means algorithm, as in Section 5.2. Solid markers indicate the representative strategies for $U100S$. Hollow markers indicate those for $U100$. This panel shows that the strategies that senders play in these two treatments are highly comparable, despite the difference in the message space. We note that the behavior in $U100S$ is less noisy than in $U100$. This can be deduced from the fact that the residual cluster, indicated by green diamonds, has a lower frequency in $U100S$ (12.9%) relative to $U100$ (21.1%). There is a higher frequency of senders who approximately best respond to receiver $U100S$ relative to $U100$. From Figures 6 and C11, we can deduce that in these treatments the best response involves a combination of blue squares and yellow triangles. These represent 63.5% and 44% of the data in $U100S$ and $U100$, respectively. This last observation is also reflected in the *average* Bayesian correlation that is induced by senders in these two treatments. We find that $\phi^B(\pi_C) = 0.41$ in $U100S$. This is significantly lower ($p < 0.01$) than the equilibrium prediction of 0.5, but higher than in $U100$ ($p < 0.05$). We conclude that senders' behavior in $U100S$ is qualitatively comparable to $U100$, but it is cleaner and less noisy than in $U100$.

[3]We conducted four sessions of $U100S$, each with 14–20 subjects (17.5 on average per session) for a total of 70 subjects. In addition to their earnings from the experiment, subjects received a \$10 show-up fee. Average earnings, including the show-up fee, were \$34 (ranging from \$14 to \$52) per session.

We now compare receivers' behavior in treatments $U100$ and $U100S$. The right panel of Figure B10 reports the average receivers' responsiveness to Bayesian posteriors belonging to four key intervals (horizontal axis). We focus attention on the posteriors induced by message $m = r$, the potentially persuasive message. The receivers' behavior in the intervals is not significantly different in the two treatments considered We conclude that receivers do not seem to react in unexpected ways to the presence of the redundant message n.

C A Closer Look at Receivers' Behavior

We take advantage of the relative simplicity of treatment $U100S$, introduced in Appendix B.2, to take a closer look at receivers' behavior. At the end of this section, we partially expand this analysis to our main treatments.

We begin by describing some aggregate features of the data in $U100S$. First, receivers' responsiveness is monotonic in the induced posterior. That is, on average, receivers are more persuaded to guess *red* by messages that carry more evidence in favor of the state being R. As highlighted in Sections 4.1.2 and 5.1, this is a robust feature of receivers' behavior that also holds in our main treatments, including $U100S$. For $U100S$, this is illustrated graphically in Figure B10 when $m = r$. When pooling message r and b, we find that, for posteriors above $\frac{1}{2}$, receivers guess *red* 57% of the time, whereas they guess *red* only 11% of the time for posteriors below $\frac{1}{2}$ ($p \leq 0.01$).

The extent of monotonicity that we observe in receivers' behavior is sufficient to confirm one of the main insights from models of communication under commitment, namely that the best response involves some degree of strategic obfuscation: an uninformative π_C is worse than a fully informative π_C, which is worse than using commitment to randomize. In Figure C11, we replicate the same exercise performed in Figure 6 for $U100S$. As was the case for $U100$ and $V100$, we find that senders' empirical expected payoff is nonmonotone in the amount of information conveyed to the receiver, in line with the theory.

Monotonicity is, of course, a mild requirement for receivers' rationality. A Bayesian receiver should choose $a = red$ for any posterior $\mu(m, \pi_C) \geq \frac{1}{2}$ and $a = blue$ otherwise. The aggregate evidence presented in Figure B10 fails to satisfy this stronger requirement of rationality. Furthermore, receivers respond to the color of the message independently of the posterior this color conveys. When $\mu(m, \pi_C) \geq \frac{1}{2}$, receivers guess $a = red$ 62% of the time if $m = r$ and 38% of the time if $m = b$. In contrast, when $\mu(m, \pi_C) < \frac{1}{2}$, receivers guess $a = red$ 21% if $m = r$ and 5% of the time if $m = b$. These differences, which are significant at the 1% level, are inconsistent with the behavior of a Bayesian receiver. Even when provided with conclusive evidence that the state is R, that is, even when $\mu(m, \pi_C) \approx 1$, some receivers nonetheless guess *blue* at least some of the time. To summarize, at the aggregate level, receivers are non-Bayesian, an observation that is in line with a large body of experimental literature (e.g., Charness and Levin, 2005; Holt, 2007, Ch. 30).

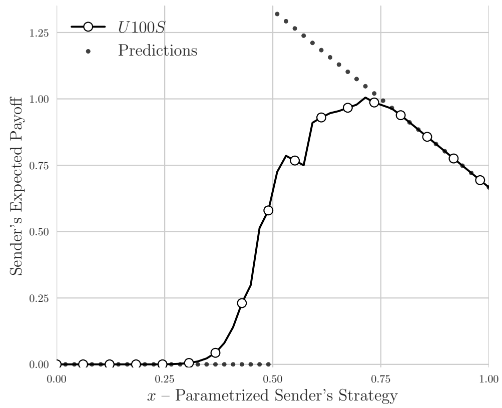


Figure C11: Probability of Guessing Red by Posterior and Message

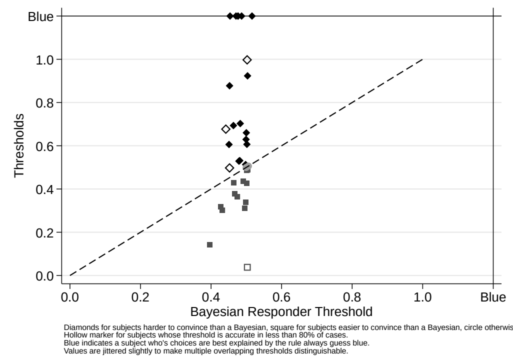


Figure C12: Estimated Thresholds: Actual Receivers vs Bayesians

To understand better whether the deviations are driven by a few subjects or shared by most, we look at individual behavior. We demonstrate that, despite not being Bayesian, receivers react to information as summarized by the posterior belief in systematic ways. In particular, we consider the possibility that subjects follow (potentially different) *threshold strategies*. A $\bar{\mu}$-threshold strategy, for $\bar{\mu} \in [0, 1]$, consists of guessing $a = red$ if and only if $\mu(m, \pi_C) \geq \bar{\mu}$. When $\bar{\mu} = \frac{1}{2}$, the receiver is Bayesian. When $\bar{\mu} \neq \frac{1}{2}$ the receiver is non-Bayesian, but behaves systematically: she requires stronger or weaker than needed evidence to choose $a = red$. Given our data, we can estimate a receiver-specific threshold that rationalizes the greatest fraction of her guesses.

The relevant data for the estimation of threshold strategies comprises pairs of induced posteriors μ and guesses a for each receiver and message. We look for a threshold $\bar{\mu} \in [0, 1]$ that minimizes $\#\{a \neq \mathbb{1}\{\mu \geq \bar{\mu}\}\}$ where a takes a value of 1 for *red* and 0 for *blue*. In other words, we find the threshold $\bar{\mu}$ that rationalizes the greatest number of choices a receiver has made.[4] We refer to the fraction of choices properly accounted for by the threshold as the *precision* of $\bar{\mu}$. Given that the sample is finite and thresholds exist on the unit interval, there will be an infinite number of thresholds with the same precision. For instance, imagine a hypothetical sample comprising only two observations: a receiver that guessed *red* given a posterior of 0.7 and guessed *blue* when the posterior was 0.4. In this case, any threshold $\bar{\mu} \in [0.4, 0.7]$ would have the same precision, namely 1. We report the midpoints of the estimated ranges.

The theory assumes receivers are Bayesian. However, notice that even a Bayesian receiver is unlikely to yield a threshold of 0.5. This is because the sample is finite. For instance, in the two-observation example proposed above, the estimated threshold is 0.55, even if the agent behaves as a Bayesian. To account for this, we compare thresholds for the receivers in our experiment with the hypothetical thresholds that we would estimate given the observed sample if the receivers were Bayesian.

[4]This method akin to *perceptrons* in machine learning; see, for instance, Murphy (2012).

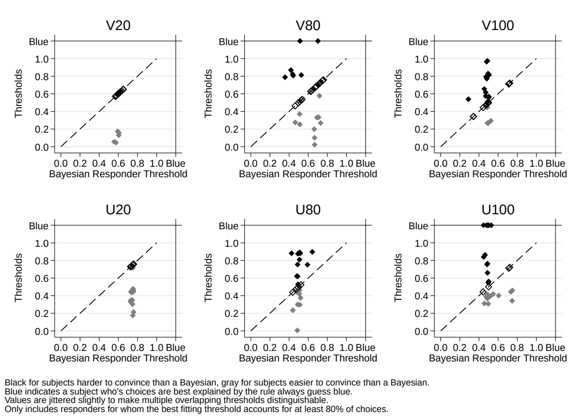


Figure C13: Estimated Threshold: Actual Receivers Against Bayesian

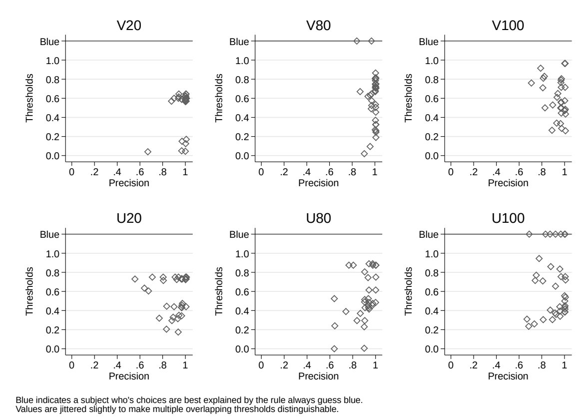


Figure C14: Estimated Threshold and Precision

Figure C12 plots the estimated threshold for each receiver (vertical axis) against those that we would have estimated from the same data if receivers were Bayesian (horizontal axis). We find that the behavior of many subjects is consistent with a threshold strategy. Almost half the receivers (46%) display behavior that is always consistent with a threshold strategy, and almost nine out of ten receivers (89%) behave consistently with a threshold strategy for more than 80% of their guesses (see Figure C16). Figure C12 reveals substantial heterogeneity in receivers' behavior. Dots lying above the 45-degree line indicate receivers who are reluctant to guess *red*, even when a Bayesian would conclude that there is enough evidence. By contrast, the points below the 45-degree line indicate subjects who are too eager to guess *red*, despite insufficient evidence from the perspective of a Bayesian agent. The aggregation of this heterogeneous behavior is partly responsible for the smoothness of aggregate responses to the posterior that is displayed in Figure B10 (right panel). Also note that Figure C12 shows a sizable fraction of receivers who exhibit behavior consistent with the Bayesian benchmark: one-quarter of the receivers have thresholds within 5 percentage points of being consistent with a Bayesian receiver; the number increases to one-third if we are more permissive and allow for a band of 10 percentage points around the Bayesian receiver.

Overall, this threshold analysis reveals three important aspects of receivers' behavior. First, the majority of receivers appear to behave in systematic ways, as summarized by threshold strategies. Second, there is substantial heterogeneity in the thresholds: some receivers are skeptical, some are approximately Bayesian, some others are gullible. Third, virtually all receivers respond to information in monotonic ways. It is thanks to this that the senders' empirical best responses (Figure C11) are qualitatively in line with the theory.

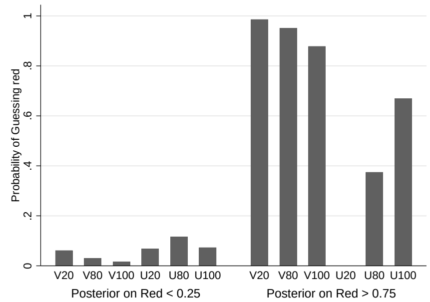


Figure C15: Frequency of $a = red$ for All Messages Given Posterior

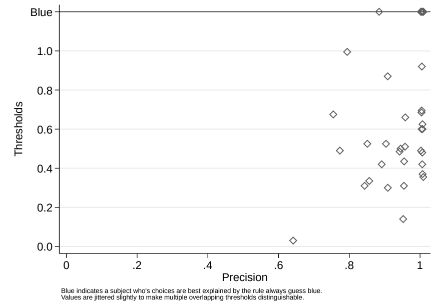


Figure C16: Estimated Threshold and Precision for Treatment U100S

C.1 Threshold Strategies in Main Treatments

Figures C13 and C14 illustrate the best-fitting thresholds and their precisions for the main treatments. Unlike for $U100S$, the main treatments feature a larger message space (three versus two). Thus, there are more choices to rationalize and achieving high precision is more difficult. Nonetheless, precision is still high: the treatment with the lowest precision still has 81% of subjects with 80% precision; across all treatments, 90% of subjects meet that criteria.

Figure C14 also shows that precision is particularly high when information is verifiable: 55% of receivers always choose in a way that is consistent with a threshold. That number is 24% for the treatments with unverifiable messages. From Figure C13, we deduce that receiver behavior is highly heterogeneous. A nontrivial fraction of subjects are close to the behavior Bayesian receivers would exhibit. There is also a substantial fraction of subjects who are skeptical, that is, they require higher-than-needed evidence to guess *red*, and there is a fraction of subjects who are instead, gullible. Finally, note that in the treatment that comes closest to the setup of a cheap talk experiment, namely $U20$, all receivers that are not compatible with the Bayesian benchmark are classified as gullible. This is in line with one of the main findings in Cai and Wang (2006). Overall, the aggregation of this heterogeneous behavior is partly responsible for the linearity of aggregate responses to the posterior that is displayed in Figure 3.[5]

Finally, in all treatments, receivers' responsiveness is monotone increasing in information. However, there are some expected differences between communication rules. As Figure C15 illustrates, in treatments with verifiable information, receivers are more likely to guess $a = red$ conditional on any message m that leads to a posterior above $3/4$. This is in part because in these treatments the frequency of extreme posteriors, that is $\mu = 1$, is higher, since information is verifiable. Conversely, the frequency of $a = red$ conditional on any message m that leads to a posterior below $1/4$

[5]This linearity may appear consistent with *probability matching*. That is, subjects guess *red* with a probability equal to the posterior belief. To test for this, we compute for each subject the mean-squared error (MSE) of the predicted guess using the estimated threshold strategies and compare it with the MSE of the probability-matching model. Across all treatments, we find that for about 84% of the receivers, threshold strategies have lower MSE than probability matching.

is lower in the verifiable treatments (it is already very low in the unverifiable treatments). Again, this is in part because the frequency of extreme posteriors, in this case $\mu = 0$, is higher in treatments with verifiable information.

D Additional Material

D.1 Remaining Proofs

Lemma 1. *Suppose information is unverifiable. Fix an arbitrary $\rho \in [0,1]$. Fix (π_C, π_R) and define $\sigma(m) = a_H$ if and only if $\mu(m, \pi_C, \pi_R) \geq q$. Then*

$$\phi(\pi_C, \pi_R, \sigma) \neq \sqrt{q\underline{\rho}} \quad \Rightarrow \quad \sum_{\theta,m} \mu_0(\theta)(\rho\pi_C(m|\theta) + (1-\rho)\pi_R(m|\theta))v(\sigma(m)) < \mu_0/q.$$

Proof. We begin by noting that, if $\sigma(m) = a_L$ for all m, then $V = 0$ and, thus, the claim holds. Therefore, suppose that there is $\emptyset \neq M' \subsetneq M$ such that $\sigma(m) = a_H$ for $m \in M'$. Fix $m' \in M'$ and $m'' \in M \setminus M'$. Let π be defined as $\pi(m'|\theta) = \sum_{m\in M'} (\rho\pi_C(m|\theta) + (1-\rho)\pi_R(m|\theta))$ and, similarly, $\pi(m''|\theta) = \sum_{m \in M\setminus M'} (\rho\pi_C(m|\theta) + (1-\rho)\pi_R(m|\theta))$, for all θ. By construction, π gives strictly positive probability to only two messages, m' and m'', inducing actions a_H and a_L, respectively. Moreover, π and (π_C, π_R) are equivalent in the sense that $\sum_{\theta,m} \mu_0(\theta)\pi(m|\theta)v(\sigma(m)) = V$ and $\phi(\pi, \sigma) = \phi(\pi_C, \pi_R, \sigma)$. Therefore, it is enough to show that $\phi(\pi, \sigma) \neq \sqrt{q\underline{\rho}}$ implies that $V < \mu_0/q$. To do so, we will argue that $V \geq \mu_0/q$ implies $\phi(\pi, \sigma) = \sqrt{q\underline{\rho}}$. Let $V \geq \mu_0/q$. Since μ_0/q is the highest achievable payoff under full commitment, it must be that $V = \mu_0/q$. To simplify notation, let $\pi_C(m'|\theta_H) = x$ and $\pi_C(m''|\theta_L) = y$. With this, $V = \mu_0 x + (1-\mu_0)(1-y) = \mu_0/q$, which can be rewritten as

$$\frac{1-\underline{\rho}}{1-q}(1-qx) = 1-y.$$

Note that since $\sigma(m') = a_H$, $\mu(m', \pi) \geq q$ or equivalently $(1-\underline{\rho})x \geq 1-y$. Together, these two equations imply that $x = 1$ and thus that $y = \underline{\rho}$. Note that these values are indeed compatible with $\sigma(m'') = a_L$, since in this case $\mu(m'', \pi) < q$. Finally, note that $\phi(\pi, \sigma)$ can be written as

$$\phi(\pi, \sigma) = \sqrt{\mu_0(1-\mu_0)}\frac{xy - (1-x)(1-y)}{\sqrt{V(1-V)}} = \sqrt{q\underline{\rho}},$$

where the last equation is obtained by substituting the values for x and y. □

Lemma 2. *Suppose information is unverifiable. Fix $\rho \in [0,1)$. For every π_C, there exists a continuation TWC equilibrium (π_R, σ, μ).*

Proof: Fix π_C and $0 \leq \rho < 1$. We consider three cases.

Case 1. Suppose that $\mu(m, \pi_C) < q$ for all m that have strictly positive probability under π_C. Define $\pi_R(\theta_H|\theta) = 1$ for all θ. Note that such π_R is trivially compatible with the TWC refinement. Moreover, define $\sigma(m) = a_L$ for all m. To complete the proof, we define $\mu(m, \pi_C, \pi_R)$ for all m. If m has zero probability under $\rho\pi_C + (1-\rho)\pi_R$, we simply let $\mu(m, \pi_C, \pi_R) = 0$. If instead m has strictly positive probability under $\rho\pi_C + (1-\rho)\pi_R$, we consider two cases. First, suppose $m \neq \theta_H$. In this case, $\pi_R(m|\theta) = 0$ for all θ, and thus $\mu(m, \pi_C, \pi_R) = \mu(m, \pi_C) < q$. Second, suppose $m = \theta_H$. To simplify notation, denote $\pi_C(\theta_H|\theta_H) = x$ and $\pi_C(\theta_H|\theta_L) = y$. Note that $\mu(\theta_H, \pi_C, \pi_R) < q$ can be rewritten as

$$(1-\underline{\rho})x - y < 0 < \frac{1-\rho}{\rho}\underline{\rho}.$$

If $x = y = 0$, the inequality holds as the left-hand-side is equal to zero. If instead $x + y > 0$, then $m = \theta_H$ has strictly positive probability under π_C. By assumption then $\mu(\theta_H, \pi_C) < q$, which implies that $(1-\underline{\rho})x - y < 0$. Therefore, $\mu(\theta_H, \pi_C, \pi_R) < q$.

Case 2. Suppose that there is a unique m' with strictly positive probability under π_C such that $\mu(m', \pi_C) \geq q$.

(i). If $m' = \theta_H$, define $\pi_R(\theta_H|\theta) = 1$ for all θ. If $\mu(\theta_H, \pi_C, \pi_R) \geq q$ define $\sigma(\theta_H) = a_H$, otherwise, define $\sigma(\theta_H) = a_L$. For all $m \neq m'$, let $\sigma(m) = a_L$. If there is $m \neq m'$ with zero probability under $\rho\pi_C + (1-\rho)\pi_R$, let $\mu(m, \pi_C, \pi_R) = 0$. We have defined a triple (π_R, σ, μ) that is a continuation TWC equilibrium given π_C.

(ii). Conversely, let $m' \neq \theta_H$. To simplify notation, let $\pi_C(\theta_H|\theta_H) = x$, $\pi_C(m'|\theta_H) = x'$, and $\pi_C(m''|\theta_H) = x''$. Similarly, let $\pi_C(\theta_H|\theta_L) = y$, $\pi_C(m'|\theta_L) = y'$, and $\pi_C(m''|\theta_L) = y''$. Clearly, $x + x' + x'' = y + y' + y'' = 1$. Define $\Lambda = (1-\underline{\rho})x - y$, $\Lambda' = (1-\underline{\rho})x' - y'$, and $\Lambda'' = (1-\underline{\rho})x'' - y''$. Note that our assumption on the interim beliefs $\mu(m, \pi_C)$ implies that $\Lambda < 0$, $\Lambda' \geq 0$, and $\Lambda'' < 0$.

- Suppose $\Lambda' \geq \frac{1-\rho}{\rho}\underline{\rho}$. Define $\pi_R(m'|\theta) = 1$ for all θ and $\sigma(m) = a_H$ if and only if $m = m'$. We have defined a triple (π_R, σ, μ) that is a continuation TWC equilibrium given π_C.
- Conversely, suppose $\Lambda' < \frac{1-\rho}{\rho}\underline{\rho}$. Define $\pi_R(\theta_H|\theta_H) = 1$, $\pi_R(\theta_H|\theta_L) = \delta$, and $\pi_R(m'|\theta_L) = 1 - \delta$. By construction, $\mu(\theta_H, \pi_C, \pi_R)$ is strictly decreasing in δ, $\mu(m', \pi_C, \pi_R)$ is strictly increasing in δ, Instead, $\mu(m'', \pi_C, \pi_R) = \mu(m'', \pi_C) < q$ and it is independent of δ. Define $\delta^* = \max\{0, \frac{\rho}{1-\rho}\Lambda + 1 - \underline{\rho}\}$ and $\delta_* = 1 - \frac{\rho}{1-\rho}\Lambda'$. Since $0 \leq \Lambda' < \frac{1-\rho}{\rho}\underline{\rho}$, $\delta_* \in [0,1]$. Similarly, since $\Lambda < 0$, $\delta^* \in [0,1]$. Suppose $\delta^* < \delta_*$. Then, let $\delta \in (\delta^*, \delta_*)$. By construction, $\mu(m, \pi_C, \pi_R) < q$ for all m. In this case, letting $\sigma(m) = a_L$ for all m concludes the proof, namely we have defined a triple (π_R, σ, μ) that is a continuation TWC equilibrium given π_C. Conversely, suppose $\delta^* \geq \delta_*$. Then, let $\delta \in [\delta_*, \delta^*]$. By construction, $\mu(m, \pi_C, \pi_R) \geq q$ for $m \in \{\theta_H, m'\}$. In this case, letting $\sigma(m) = a_L$ if and only if $m = m''$ concludes the proof.

Case 3. Finally, we consider the case in which there are exactly two messages with strictly positive probability under π_C such that $\mu(m', \pi_C) \geq q$. Denote the set of such messages $\bar{M} \subsetneq M$.

(i). Suppose $\theta_H \in \bar{M}$. Without loss of generality, let $m' \in \bar{M}$. By the martingale property, $\mu(m'', \pi_C) < q$ for $m'' \in M \setminus \bar{M}$. Define $\pi_R(\theta_H|\theta_H) = 1$, $\pi_R(\theta_H|\theta_H) = \delta$, and $\pi_R(\theta_H|\theta_L) = 1 - \delta$. As for Case 2, $\mu(\theta_H, \pi_C, \pi_R)$ is strictly decreasing in δ, while $\mu(m', \pi_C, \pi_R)$ is strictly increasing in δ. Instead, $\mu(m'', \pi_C, \pi_R) = \mu(m'', \pi_C) < q$ and independent of δ. Moreover, since by assumption $\mu(\theta_H, \pi_C) \geq q$, if $\delta = 0$, $\mu(\theta_H, \pi_C, \pi_R) \geq q$. Similarly, since by assumption $\mu(m', \pi_C) \geq q$, if $\delta = 1$, $\mu(\theta_H, \pi_C, \pi_R) \geq q$. Let δ^* be the unique δ such that $\mu(\theta_H, \pi_C, \pi_R) = q$. Similarly, let δ_* be the unique δ such that $\mu(\theta_H, \pi_C, \pi_R) = q$. Suppose $\delta^* < \delta_*$. Then, let $\delta \in (\delta^*, \delta_*)$. By construction, $\mu(m, \pi_C, \pi_R) < q$ for all $m \in \bar{M} = \{\theta_H, m'\}$. In this case, letting $\sigma(m) = a_L$ for all m concludes the proof. Conversely, suppose $\delta^* \geq \delta_*$. Then, let $\delta \in [\delta_*, \delta^*]$. By construction, $\mu(m, \pi_C, \pi_R) \geq q$ for $m \in \bar{M} = \{\theta_H, m'\}$. In this case, letting $\sigma(m) = a_L$ if and only if $m = m''$ concludes the proof.

(ii). Finally, suppose that $\theta_H \notin \bar{M} = \{m', m''\}$. We first consider a simpler problem, in which m' and m'' are treated as a single message, labeled $\bar{m}$. To this purpose, define $\bar{\pi}_C(\bar{m}|\theta) = \pi_C(m'|\theta) + \pi_C(m''|\theta)$ and $\bar{\pi}_C(\theta_H|\theta) = \pi_C(\theta_H|\theta)$ for all θ. Define $\bar{\pi}_R(\theta_H|\theta_H) = 1$, $\bar{\pi}_R(\theta_H|\theta_L) = \delta$, and $\bar{\pi}_R(\bar{m}|\theta_L) = 1 - \delta$. Our goal is to find $\bar{\delta}$ such that $\mu(m, \bar{\pi}_C, \bar{\pi}_R) < q$ for $m \in \{\theta_H, \bar{m}\}$. These two inequalities are equivalent to

$$\frac{\rho}{1-\rho}\Big((1-\underline{\rho})x - y\Big) + 1 - \underline{\rho} < \delta \qquad \text{and} \qquad \delta < 1 - \frac{\rho}{1-\rho}\Big((1-\underline{\rho})\bar{x} - \bar{y}\Big),$$

respectively. Therefore, such a $\bar{\delta}$ exists if $\frac{\rho}{1-\rho}((1-\underline{\rho})(x+\bar{x}) - (y - \bar{y})) < \underline{\rho}$, which always holds (recall that, by construction, $x + \bar{x} = 1 = y + \bar{y}$). To complete the proof, we now define $\pi_R(\theta_H|\theta) = \bar{\pi}_R(\theta_H|\theta)$, $\pi_R(m'|\theta_L) = \alpha(1-\bar{\delta})$, and $\pi_R(m''|\theta_L) = (1-\alpha)(1-\bar{\delta})$. Our goal is to find a $\bar{\alpha} \in [0,1]$ such that $\mu(m, \pi_C, \pi_R) < q$ for $m \in \{m', m''\}$. Begin by noting that:

$$1 - \bar{\delta} > \frac{\rho}{1-\rho}\Big((1-\underline{\rho})\bar{x} - \bar{y}\Big) = \underbrace{\frac{\rho}{1-\rho}\Big((1-\underline{\rho})x' - y'\Big)}_{A \geq 0} + \underbrace{\frac{\rho}{1-\rho}\Big((1-\underline{\rho})x'' - y''\Big)}_{B \geq 0} = A + B.$$

Also, note that $\mu(m', \pi_C, \pi_R) < q$ iff $A < \alpha(1-\bar{\delta})$. Similarly, $\mu(m', \pi_C, \pi_R) < q$ iff $B < (1-\alpha)(1-\bar{\delta})$. To find $\bar{\alpha}$, define $g(\alpha) = \alpha(1-\bar{\delta}) - A$ and $f(\alpha) = (1-\alpha)(1-\bar{\delta}) - B$ and let $\bar{\alpha}$ be the unique solution to $g(\alpha) = f(\alpha)$, namely that is $\bar{\alpha} = \frac{(1-\bar{\delta})+A-B}{2(1-\bar{\delta})}$. Since $A, B \geq 0$ and $A + B < 1 - \bar{\delta}$, then $A < 1 - \bar{\delta}$ and $B < 1 - \bar{\delta}$. This implies that $\bar{\alpha} \in [0,1]$. Finally, note that $g(\bar{\alpha}) = f(\bar{\alpha}) > 0$, implying that $\mu(m, \pi_C, \pi_R) < q$ for $m \in \{m', m''\}$. □

Proof of Proposition 3. Assume that information is unverifiable. Fix $q' > q > \mu_0$. Consider $\rho \geq \underline{\rho}' := \frac{q'-\mu_0}{q'(1-\mu_0)}$. Since $q' > q$, $\underline{\rho}' > \underline{\rho} := \frac{q-\mu_0}{q(1-\mu_0)}$ and, thus, $\rho \geq \underline{\rho}$ as well. By Theorem 1, all equilibria when the persuasion threshold is q' are FCC, namely they induce correlation $\sqrt{q'\underline{\rho}}$. Similarly, all equilibria when the persuasion threshold is q are FCC, namely they induce correlation $\sqrt{q\underline{\rho}}$. Since $q' > q$, the equilibrium correlation induced when the persuasion threshold is q' is higher than that induced when the persuasion threshold is q. $\square$

D.2 Correlation and Blackwell Informativeness

D.2.1 The Informativeness of an Outcome

Fix $\mu_0 \in (0,1)$, $\rho \in [0,1]$, and Π. Fix strategies (π_C, π_R, σ). Let the outcome induced by (π_C, π_R, σ) be the function $\eta : \Theta \to \Delta(A)$, defined as $\eta(a|\theta) = \sum_m (\rho\pi_C(m|\theta) + (1-\rho)\pi_R(m|\theta))\sigma(a|m)$, for all a and θ. We can think of an outcome η as an information structure on its own, which could be informative about θ. It is as if an external observer were to learn about θ only by observing the action a taken by the receiver. Say that an outcome η' is Blackwell more-informative than η if there is a garbling $g : A \to \Delta(A)$ such that $\eta(a|\theta) = \sum_{a'} g(a|a')\eta'(a'|\theta)$ for all a and θ. The next result shows that the correlation ϕ is a completion of the Blackwell order on the space of outcomes.

Remark 1. *Let (π_C, π_R, σ) and $(\pi'_C, \pi'_R, \sigma')$ be two strategy profiles and η and η' their respective outcomes. Suppose that η' is Blackwell more-informative than η. Then, $\phi(\pi'_C, \pi'_R, \sigma') \geq \phi(\pi_C, \pi_R, \sigma)$.*

Proof: Let η be the outcome induced by (π_C, π_R, σ). To simplify notation, define $\alpha = \eta(a_H|\theta_H)$ and $\beta = \eta(a_H|\theta_L)$. The correlation is equal to

$$\phi(\pi_C, \pi_R, \sigma) = \frac{\sqrt{\mu_0(1-\mu_0)}}{\sqrt{(\mu_0\alpha + (1-\mu_0)\beta)(1-\mu_0\alpha-(1-\mu_0)\beta)}}(\alpha-\beta).$$

Consider an external observer with prior belief μ_0 that the state is θ_H. She observes the realized action a from η. The distribution of the observer's posterior belief is:

$$\mu(\theta_H|a) = \begin{cases} \dfrac{\mu_0\alpha}{\mu_0\alpha + (1-\mu_0)\beta} & \text{with prob. } \Pr(a_H) = \mu_0\alpha + (1-\mu_0)\beta \\ \dfrac{\mu_0(1-\alpha)}{\mu_0(1-\alpha) + (1-\mu_0)(1-\beta)} & \text{with prob. } Pr(a_L) = \mu_0(1-\alpha) + (1-\mu_0)(1-\beta) \end{cases}$$

The variance of such distribution is:

$$\begin{aligned} \mathbb{V}_{a\sim\eta}(\mu(\theta_H|a)) &= \mathbb{E}_{a\sim\eta}(\mu(\theta_H|a)^2) - \mathbb{E}_{a\sim\eta}(\mu(\theta_H|a))^2 \\ &= \mathbb{E}_{a\sim\eta}(\mu(\theta_H|a)^2) - \mu(\theta_H)^2 \\ &= \mu(\theta_H)^2\Big(\frac{\alpha^2}{\mu(\theta_H)\alpha + \mu(\theta_L)\beta} + \frac{(1-\alpha)^2}{1-\mu(\theta_H)\alpha - \mu(\theta_L)\beta} - 1\Big) \\ &= \frac{\mu(\theta_H)^2\mu(\theta_L)^2}{(\mu_0\alpha + (1-\mu_0)\beta)(1-\mu_0\alpha-(1-\mu_0)\beta)}(\alpha-\beta)^2, \end{aligned}$$

where we used the fact that $\mathbb{E}_{a\sim\eta}(\mu(\theta_H|a)) = \mu_0$, by the martingale property. Therefore, we have established that

$$\phi(\pi_C, \pi_R, \sigma) = \sqrt{\frac{\mathbb{V}_{a\sim\eta}(\mu(\theta_H|a))}{\mu_0(1-\mu_0)}}.$$

That is, for any μ_0 and (π_C, π_R, σ), the state-action correlation ϕ is proportional to the standard deviation of the distribution of the implied posterior beliefs.

We can now prove the claim. Fix outcomes η' and η. By Blackwell and Girshick (1979, Theorem 12.2.2), η' is Blackwell more informative than η if and only if, for all convex functions $f : \Delta(\Theta) \to \mathbb{R}$,

$$\mathbb{E}_{a\sim\eta'}\Big(f(\mu(\theta_H|a))\Big) \geq \mathbb{E}_{a\sim\eta}\Big(f(\mu(\theta_H|a))\Big).$$

Note that, in particular, $f(\mu(\theta_H|a) = (\mu(\theta_H|a) - \mu(\theta_H))^2$ is convex and that

$$\mathbb{E}_{a\sim\eta}\Big(f(\mu(\theta_H|a))\Big) = \mathbb{V}_{a\sim\eta}(\mu(\theta_H|a)).$$

Therefore, if η' is Blackwell more informative than η, then

$$\mathbb{V}_{a\sim\eta'}(\mu(\theta_H|a)) \geq \mathbb{V}_{a\sim\eta}(\mu(\theta_H|a)) \quad \Rightarrow \quad \sqrt{\frac{\mathbb{V}_{a\sim\eta'}(\mu(\theta_H|a))}{\mu_0(1-\mu_0)}} \geq \sqrt{\frac{\mathbb{V}_{a\sim\eta}(\mu(\theta_H|a))}{\mu_0(1-\mu_0)}},$$

which implies that $\phi(\pi'_C, \pi'_R, \sigma') \geq \phi(\pi_C, \pi_R, \sigma)$. □

D.2.2 The Informativeness of a Sender's Strategy

In the paper, we distinguish between the information "sent" by the sender and the information "received" by the receiver. The latter is measured by ϕ and must inevitably rely on the entire outcome η, which combines the observed strategies of both sender and receiver. To measure information "sent," instead, there are at least two natural directions, which we are both explored in the paper and give results that are qualitatively similar.

The first approach is to use ϕ^B, the informativeness of the hypothetical outcome induced by the sender's strategy and that of a Bayesian receiver who best responds to it. It is immediate to see that Remark 1 extends to the Bayesian correlation ϕ^B. More specifically, we can show that the correlation measure ϕ^B is a completion of the Blackwell order on the space of outcomes that are induced by a strategy profile (π, σ^B).

The second approach consists of using the variance of the distribution of Bayesian posteriors that are induced by the sender's strategy. In the next remark, we show that this alternative measure of information "sent" is proportional to the posterior divergence ψ^B, which we used in Section 4.2. Fix μ_0 and a sender's strategy $\pi : \Theta \to \Delta(M)$. Strategy $\pi \in \Pi$ can indicate a commitment-stage strategy, a revision-stage strategy, or a mixture of the two. To simplify notation, denote by $\mu(m)$ the posterior belief that $\theta = \theta_H$ conditional on observing message m under π.[6] Recall that the posterior divergence is defined as $\psi^B(\pi) = \mathbb{E}_{m\sim\pi}(\mu(m)|\theta_H) - \mathbb{E}_{m\sim\pi}(\mu(m)|\theta_L)$. The next result shows that ψ^B is a completion of the Blackwell order on the space of strategies π. To do so, the proof illustrates that $\psi^B(\pi)$ is proportional to the variance of the distribution of the posterior beliefs induced by π.

Remark 2. *Let $\pi, \pi' : \Theta \to \Delta(M)$. Suppose that π' is Blackwell more informative than π. That is, suppose there exists a garbling $g : M \to \Delta(M)$ such that $\pi(m|\theta) = \sum_{m'} g(m|m')\pi'(m'|\theta)$ for all m and θ. Then $\psi^B(\pi') \geq \psi^B(\pi)$.*

Proof. Let $\mu_0 \in (0, 1)$. We rewrite $\psi^B(\pi)$ as a convex function of posteriors $\mu(m)$:

$$\begin{aligned}
\psi^B(\pi) &= \mathbb{E}_m(\mu(m)|\theta_H) - \mathbb{E}_m(\mu(m)|\theta_L) \\
&= \sum_m \mu(m)\pi(m|\theta_H) - \sum_m \mu(m)\pi(m|\theta_L) \\
&= \sum_m \mu(m)\Big(\pi(m|\theta_H) - \pi(m|\theta_L)\Big) \\
&= \sum_m \mu(m)\Big(\frac{\pi(m|\theta_H)}{\Pr_\pi(m)} - \frac{\pi(m|\theta_L)}{\Pr_\pi(m)}\Big)\Pr_\pi(m) \\
&= \sum_m \mu(m)\Big(\frac{\mu(m)}{\mu_0} - \frac{1-\mu(m)}{(1-\mu_0)}\Big)\Pr_\pi(m) \\
&= \sum_m \frac{\mu(m)^2 - \mu(m)\mu_0}{\mu_0(1-\mu_0)}\Pr_\pi(m) \\
&= \frac{\mathbb{V}_{m\sim\pi}(\mu(m))}{\mu_0(1-\mu_0)}.
\end{aligned}$$

The variance $\mathbb{V}_{m\sim\pi}(\mu(m))$ is a convex function $\mu(m)$. By Blackwell and Girshick (1979, Theorem 12.2.2), if π' is Blackwell more-informative than π, $\psi^B(\pi') \geq \psi^B(\pi)$. □

These results indicate that both ϕ^B and ψ^B are valid ways to quantify the amount of information sent by senders. In Section 4.2, we discuss both measures and argue that they lead to qualitatively similar conclusions. It is useful to discuss their similarities and differences. First, ϕ^B can be directly compared to ϕ, while ψ^B cannot. In the data, we find that the average $\phi - \phi^B$ is negative, suggesting that receivers further garble the information they have received. Second, ϕ^B exploits the fact that we know u, whereas ψ^B is "utility-free." This is important because not *all* information is useful to our receivers. Let us consider an example. Fix $\mu_0 = 1/3$ and $q = 1/2$. Let π be uninformative, in the sense that $\mu(m) = \mu_0$ for all m. Let π' induce posterior $\mu(m) = 2/5$ with probability $5/6$ and posterior is $\mu(m) = 0$ with remaining probability. None of these strategies can change the receiver's behavior, since $q > \mu(m)$ for all m. Clearly, π' is Blackwell more informative than π. Both ψ^B and ϕ^B agree with this order. However, $\psi^B(\pi') > \psi^B(\pi)$ whereas $\phi^B(\pi') = \phi^B(\pi)$. The reason for this is that π' does not contain information that is more useful to *our* receivers than π.

[6]Without loss of generality, let $\mu(m) = 0$ if m has zero probability under π.

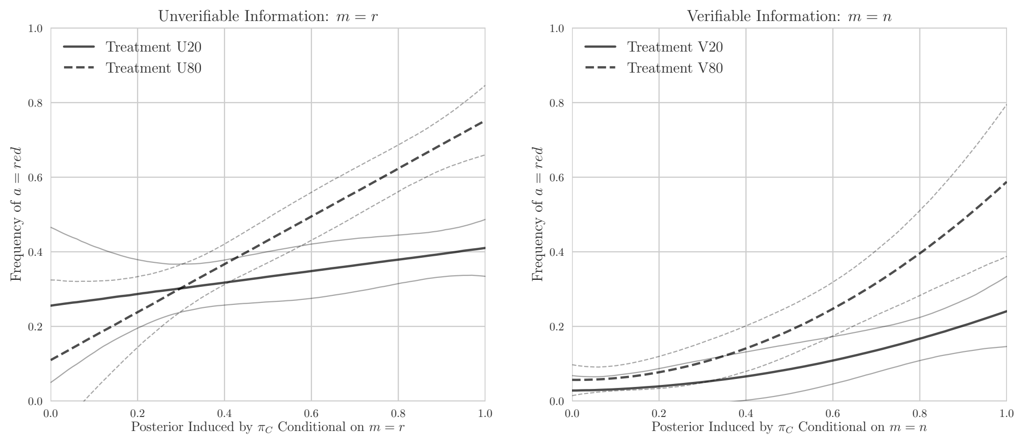


Figure D17: Receiver's Response to Persuasive Messages: ρ = 0.2 vs. ρ = 0.80

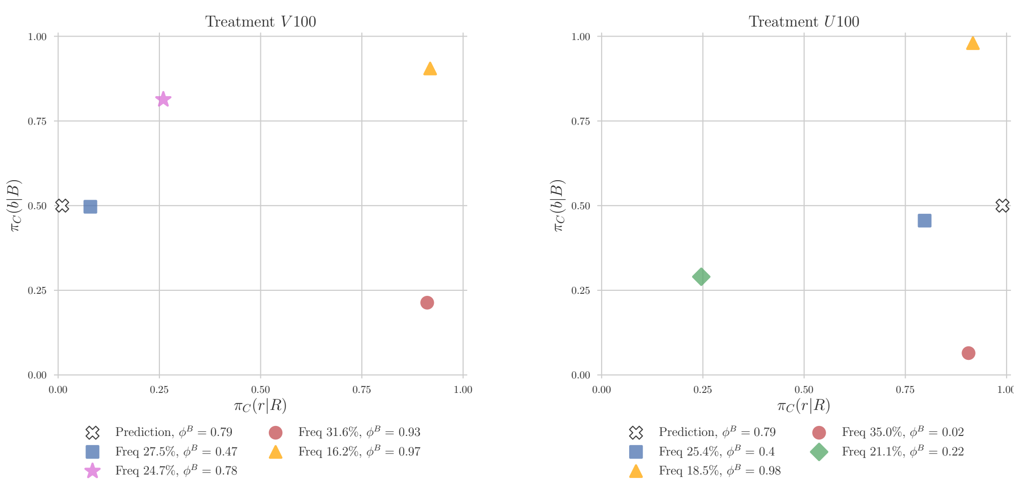


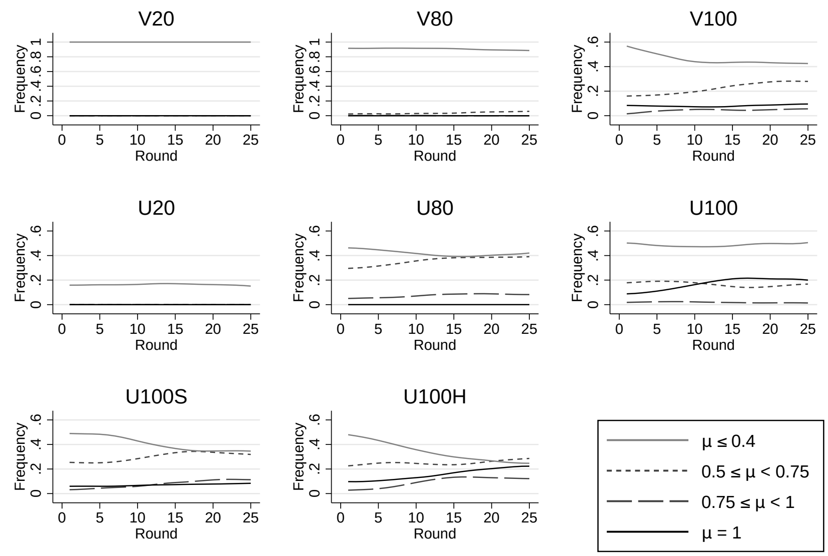
Figure D18: k-Means – Representative Strategies in Treatments with Full Commitment

Table D7: p-Values of Statistical Tests

Model / Subject / Session / Bootstrap	Linear / RE / Cluster	Linear / RE / RE	Pr(T)obit / RE / Cluster	Pr(T)obit / RE / RE	Linear / FE / Cluster / CATs	Linear / FE / Cluster
Test						
Left panel Figure 2, all bars = 0 when ball is R	0.000	0.000				
Left panel Figure 2, all bars = 0 when ball is B	0.000	0.000				
Right panel Figure 2, r message bar = 0 when ball is R	0.000	0.000				
$\phi_C^B = \phi_R^B$ in U80	0.000	0.000	0.000	0.996		
$\phi_C^B = \phi_R^B$ in V80	0.000	0.000	0.006	0.000		
$\Pr\left(red \vert m = r, \mu < \frac{1}{2}\right) = \Pr\left(red \vert m = r, \mu \geq \frac{1}{2}\right)$ in U20	0.053	0.002	0.083	0.004	0.150	0.126
$\Pr\left(red \vert m = r, \mu < \frac{1}{2}\right) = \Pr\left(red \vert m = r, \mu \geq \frac{1}{2}\right)$ in U100	0.000	0.000	0.024	0.000	0.040	0.021
$\Pr\left(red \vert m = r, \mu < \frac{1}{2}, U20\right) = \Pr\left(red \vert m = r, \mu < \frac{1}{2}, U100\right)$	0.627	0.535	0.718	0.610		
$\Pr\left(red \vert m = r, \mu \geq \frac{1}{2}, U20\right) = \Pr\left(red \vert m = r, \mu \geq \frac{1}{2}, U100\right)$	0.000	0.001	0.002	0.003		
$\Pr\left(red \vert m = n, \mu < \frac{1}{2}\right) = \Pr\left(red \vert m = n, \mu \geq \frac{1}{2}\right)$ in V20	0.038	0.002	0.133	0.006	0.257	0.163
$\Pr\left(red \vert m = n, \mu < \frac{1}{2}\right) = \Pr\left(red \vert m = n, \mu \geq \frac{1}{2}\right)$ in V100	0.000	0.000	0.000	0.000	0.022	0.014
$\Pr\left(red \vert m = r, \mu < \frac{1}{2}, V20\right) = \Pr\left(red \vert m = r, \mu < \frac{1}{2}, V100\right)$	0.566	0.674	0.536	0.452		
$\Pr\left(red \vert m = r, \mu \geq \frac{1}{2}, V20\right) = \Pr\left(red \vert m = r, \mu \geq \frac{1}{2}, V100\right)$	0.000	0.000	0.000	0.000		
$\phi(V20) = \phi(V80)$	0.217	0.215				
$\phi(V80) = \phi(V100)$	0.001	0.020	0.258	0.451		
$\phi(U20) = \phi(U80)$	0.002	0.001				
$\phi(U80) = \phi(U100)$	0.696	0.676	0.486	0.441		
$\phi(V20) = \phi(U20)$	0.000	0.000				
$\phi(V80) = \phi(U80)$	0.000	0.000				
$\phi(V100) = \phi(U100)$	0.000	0.000	0.000	0.000		
$\phi^B(V20) = \phi^B(V80)$	0.156	0.130				
$\phi^B(V80) = \phi^B(V100)$	0.032	0.052	0.608	0.648		
$\phi^B(U20) = \phi^B(U80)$	0.000	0.000				
$\phi^B(U80) = \phi^B(U100)$	0.957	0.925	0.711	0.661		
$\phi^B(V20) = \phi^B(U20)$	0.000	0.000				
$\phi^B(V80) = \phi^B(U80)$	0.000	0.000				
$\phi^B(V100) = \phi^B(U100)$	0.000	0.000	0.000	0.000		

D.3 Statistical Tests

The p-values reported in the main text are obtained by regressing the variable of interest on the relevant regressor (sometimes an indicator variable) with subject-level random effects and clustering of the variance-covariance matrix at the session level. This specification has the advantage of being uniform (the same throughout the paper), it directly accounts for heterogeneity across subjects via the random effects (as the paper documents, there is clear evidence of heterogeneity between subjects), and it permits unmodeled dependencies between observations from the same session (see Fréchette, 2012, where such possibilities are discussed). However, it does not directly account for the fact that we are often dealing with a limited dependent variable. Also, clustering with a small number of clusters can lead to insufficient corrections (see Cameron and Miller, 2015, for a survey). But this observation relies mostly on simulations that do not necessarily mirror the situation of most laboratory experiments. In particular, the extent of the problem is found to depend on the size of the within-session correlation (see, for example, Carter et al., 2017). For many experiments, such correlation can be expected to be low (once the appropriate factors are controlled for). Hence, we are more concerned with controlling for the source of dependencies across the observations of a given subject than for the within-session correlations (see also Appendix A.4 of Embrey et al. (2017) for a discussion of these issues).



Posterior following a critical message: no message for V treatments and red message for U treatments
V20 and V80 are drawn with a different y-axis.

Figure D19: Senders' Frequency of Inducing $\mu(m, \pi_C, \pi_R)$

In Table D7 we document the robustness of the tests reported in the text by exploring alternative specifications. These include directly accounting for the limited nature of the dependent variable by using a probit or Tobit when appropriate. When possible we also report bootstrapped estimates that have been shown to perform better when the number of clusters is small (cluster-adjusted t-statistics or CAT) and that allow for subject-specific fixed-effects (Ibragimov and Müller, 2010). When we report those we also include results from a standard subject specific fixed-effects estimation with session clustering to provide a benchmark. As can be seen, p-values are not systematically larger for CATs than with the "standard" clustering, nor are they very different when estimating a probit or tobit.[7] As a whole, results are fairly robust: out of the 28 hypotheses tested, for only five of them are results not the same for all tests reported (in the sense of being consistently significant–or not–at the 10% level). The few cases in which there are differences are for the most part not difficult to make sense of. Two of them involve comparing $V80$ and $V100$, where the difference is small in magnitude. Hence, whether or not the difference is statistically significant is not clear, but either way it is not large. In most other cases, the p-values are either under the 0.1 cutoff or just slightly above.

D.4 Subjects' Behavior Over Time

Figures D19 and D20 illustrate changes in behavior over the course of the experiment.

Senders. Figure D19 studies senders by coarsely separating their strategies by the posterior they

[7]Note that if a tobit could have been estimated but is not reported, it means that the dependant variable was not actually censored.

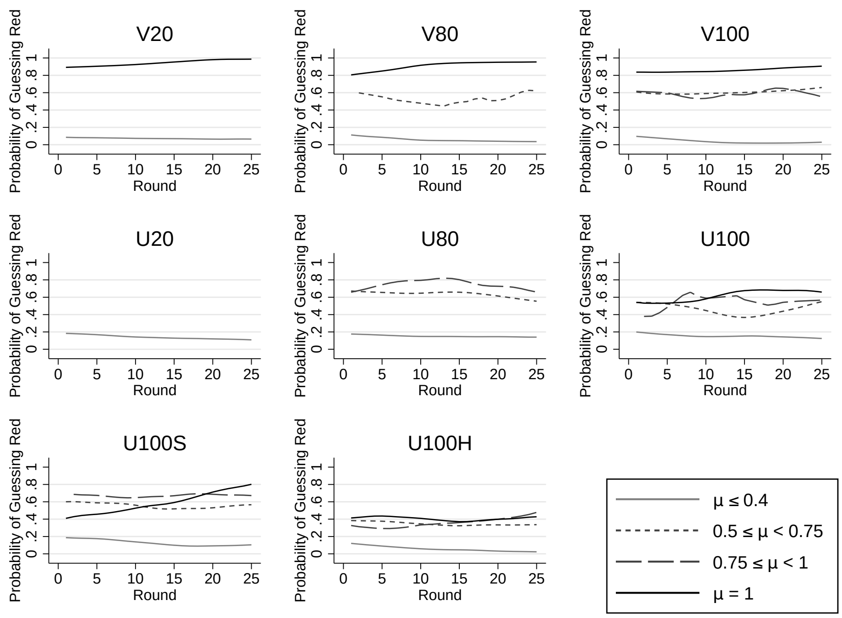


Figure D20: Receivers's Frequency of a = *red* Grouped by Posterior (μ)

induce conditional on the *persuasive* message; that is, message n under verifiable information and r otherwise. Four posterior intervals are considered: low ($\mu < 0.4$), close to full-commitment equilibrium ($0.5 \geq \mu < 0.75$), high ($0.75 \leq \mu < 1$), and maximal ($\mu = 1$). We excluded posteriors in the interval $0.4 < \mu < 0.5$. As the figure shows, overall there are very few changes over time (at least, no change across these groups of posteriors). Notable exceptions are treatment $U100H$ and, to a lesser extent $U100S$ and $U100$, where senders seem to learn to provide more information over time.

Receivers. Figure D20 studies receivers and displays changes in terms of the likelihood a given posterior leads to a a = *red*. Overall, time effects are limited. There appears to be an increase in the frequency of a = *red* conditional on higher posteriors ($U80$ is one exception) and a decrease of such frequency conditional on lower posteriors.

D.5 $V0$ and $U0$

In Table D8, we report the average revision-stage strategies π_R, for treatments $U20$ and $V20$. This stage of these treatments represents the closest point in our data to the hypothetical treatments $U0$ and $V0$. For $U20$, the table shows that the average revision strategy is akin to babbling. In particular, all messages lead to a posterior belief that is well below the persuasion threshold $q = 1/2$ (recall that in the experiment the prior is $\mu_0 = 1/3$). Therefore, following each message, a Bayesian receiver would always guess *blue*. For $V20$, the same table shows that the R-type sender almost always sends message r, while the B-type sender mostly sends message n. Given this, a Bayesian receiver would almost fully learn the state. In other words, unraveling would happen most of the time.

Table D8: Average Revision-Stage Strategies in $U20$ and $V20$

	$U20$				$V20$		
π_R	$m = \theta_H$	$m = \theta_L$	$m = n$	π_R	$m = \theta_H$	$m = \theta_L$	$m = n$
θ_H	.89	.06	.05	θ_H	.92	0	.08
θ_L	.64	.24	.12	θ_L	0	.25	.75

D.6 Receivers' Behavior and Revealed Information

In this section, we apply methods from Caplin and Martin (2021) to study whether the receivers' behavior reveals that they are indeed better informed in $U100$ vs $U20$. We observe the behavior of receivers who take guesses upon receiving information from two different experiments, labeled E_{20} and E_{100}. Is the receiver more informed under one or the other experiment? The answer to this question is trivial if we know the utility of the receiver and which experiments she observed. In our setting, these are all details of the problem that we know. However, in this appendix, we will assume that we do not know what the "true" utility function of the receiver is. Instead, let us assume that the receiver earns an unknown payoff $u(x_r) \in \mathbb{R}$, when correctly guessing that the state is R, that she earns $u(x_b) \in \mathbb{R}$ when correctly guessing that the state is B, and that she earns $u(x_0) \in \mathbb{R}$ when guessing incorrectly. Note that we allow $u(x_r)$, $u(x_b)$, and $u(x_0)$ to be positive or negative. Similarly, we may not know how the receivers truly understand the experiments E_{20} and E_{100}. Thus, we assume that we do not observe them.

Because the space of strategies is extremely large, we will focus attention on the subset of commitment strategies that satisfies $\pi_C(r|R) \geq .95$ and $\pi_C(b|B) \geq .95$. We do not know what the receiver understands from these strategies, whether she misinterprets them entirely, or how this depends on the treatment. This is what we seek to study.[8]

For each treatment, we observe a state-dependent stochastic choice (SDSC) dataset, which consists of a large number of guesses, $a \in \{red, blue\}$, taken by the receiver conditional on the state, $\theta \in \{R, B\}$. Such a dataset can be summarized in a matrix $P_i = (P_i(a, \theta))_{a \in A, \theta \in \Theta}$ where $i \in \{20, 100\}$. Based on the comparison between P_{20} and P_{100}, we would like to conclude that the receiver is "revealed to be more informed" under E_{100} rather than E_{20}, consistent with our conclusion from Section 4. In Table D9, we report P_{20} and P_{100} computed from our treatments $U20$ and $U100$.

Without loss of generality, we can normalize one of the unknowns, so let $u(x_0) = 0$. Following Caplin and Martin (2021), we can use NIAS (No Improving Action Switches) inequalities to find the set of utilities u for which there are experiments consistent with P_{20} and P_{100}. This amounts

[8] Our conclusion in this exercise is unchanged if we study the receiver's behavior unconditional on π_C.

Table D9

	$U20$			$U100$	
P_{20}	$a = Red$	$a = Blue$	P_{100}	$a = Red$	$a = Blue$
$\theta = R$	.13	.20	$\theta = R$	.25	.08
$\theta = B$	.13	.54	$\theta = B$	.04	.63

to finding the set of utilities $(u(x_r), u(x_r)) \in \mathbb{R}^2$ such that, for all $i \in \{20, 100\}$, and for all $a, a' \in \{Red, Blue\}$, the following inequality is satisfied:

$$P_i(a, R)u(x(a, R)) + P_i(a, B)u(x(a, B)) \geq P_i(a, R)u(x(a', R)) + P_i(a, B)u(x(a', B)).$$

In the formula above, we defined $x(Red, R) = x_r$, $x(Blue, B) = x_b$, and x_0 otherwise. These four NIAS inequalities lead to the following system:

$$\begin{cases} u(x_r) \geq \frac{4}{25}u(x_b) \\ u(x_r) \leq \frac{63}{8}u(x_b) \\ u(x_r) \geq u(x_b) \\ u(x_r) \leq \frac{54}{20}u(x_b) \end{cases}$$

whose set of solutions is: $\{u(x_r), u(x_b) \in \mathbb{R}^2_+ : u(x_b) \leq u(x_r) \leq \frac{54}{20}u(x_b)\}$. Note that all utilities consistent with NIAS satisfy $u(x_r) \geq 0$ and $u(x_b) \geq 0$. Therefore, we can conclude that:

$$\sum_{\theta,a} P_{100}(a, \theta)u(x(a, \theta)) \geq \sum_{\theta,a} P_{20}(a, \theta)u(x(a, \theta)).$$

In other words, the value of information in $U100$ is higher than that in $U20$. This shows that receivers are revealed to be on average more informed under E_{100} rather than E_{20}, corroborating our evidence from Section 4.1.2.

D.7 Gaussian Mixture Model

The k-means algorithm does not allow for confidence intervals. One may wonder how confidently each observation is assigned to its cluster. To answer this question, we estimated a Gaussian mixture model (GMM) in which the centroid of each cluster is given and computed with k-means (i.e., they are those in Figures 7 and 8) while the variance of each cluster is estimated from the data. That is, we estimate a GMM with a single parameter for the variance of the errors. With this model, we can compute the posterior probabilities of each assignment, which capture how confidently we can assign an observation to its cluster.

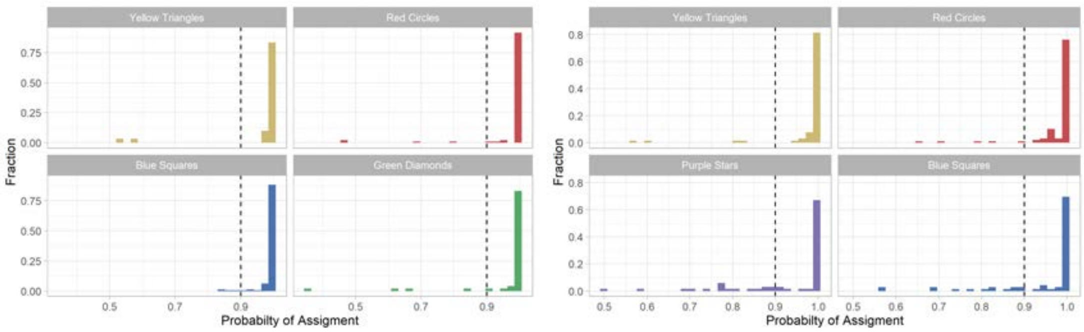


Figure D21: Posteriors Probabilities of k-Means Assignments for $U80$ (left panel) and $V80$ (right panel)

Figure D21 plots the posterior assignments of the clusters computed in that fashion for treatments $U80$ (left panel) and $V80$ (right panel) As can be seen, the posterior for the vast majority of observed strategies is extremely high. Note that, in each of the eight clusters, at least three-quarters of the strategies are classified with a posterior that is above 90%; and for six of the eight clusters that is true for more than 90% of the strategies. In fact, for half of the clusters less than 5% of the strategies are classified with a probability below 90%. This exercise shows that the cluster assignment from Section 5.2 is quite robust.

E Design

E.1 Graphical Interface

Figures E22 and E23 shows the software interface of our experiment. More specifically, Figures E22 show the commitment, revision, and guessing stages. To avoid any possible framing, the experiment referred to the first two with more neutral labels, “Communication” and “Update.” Figure E23 shows the feedback screen, where all relevant information is reported to both players.

E.2 Sample Instructions

In this section, we reproduce instructions for one of our treatments, V80. These instructions were read out aloud so that everybody could hear. A copy of these instructions was handed out to the subject and available at any point during the experiment. Finally, while reading these instructions, screenshots similar to those in Figures E22 and E23 were shown with a projector to ease the exposition and the understanding of the tasks.

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

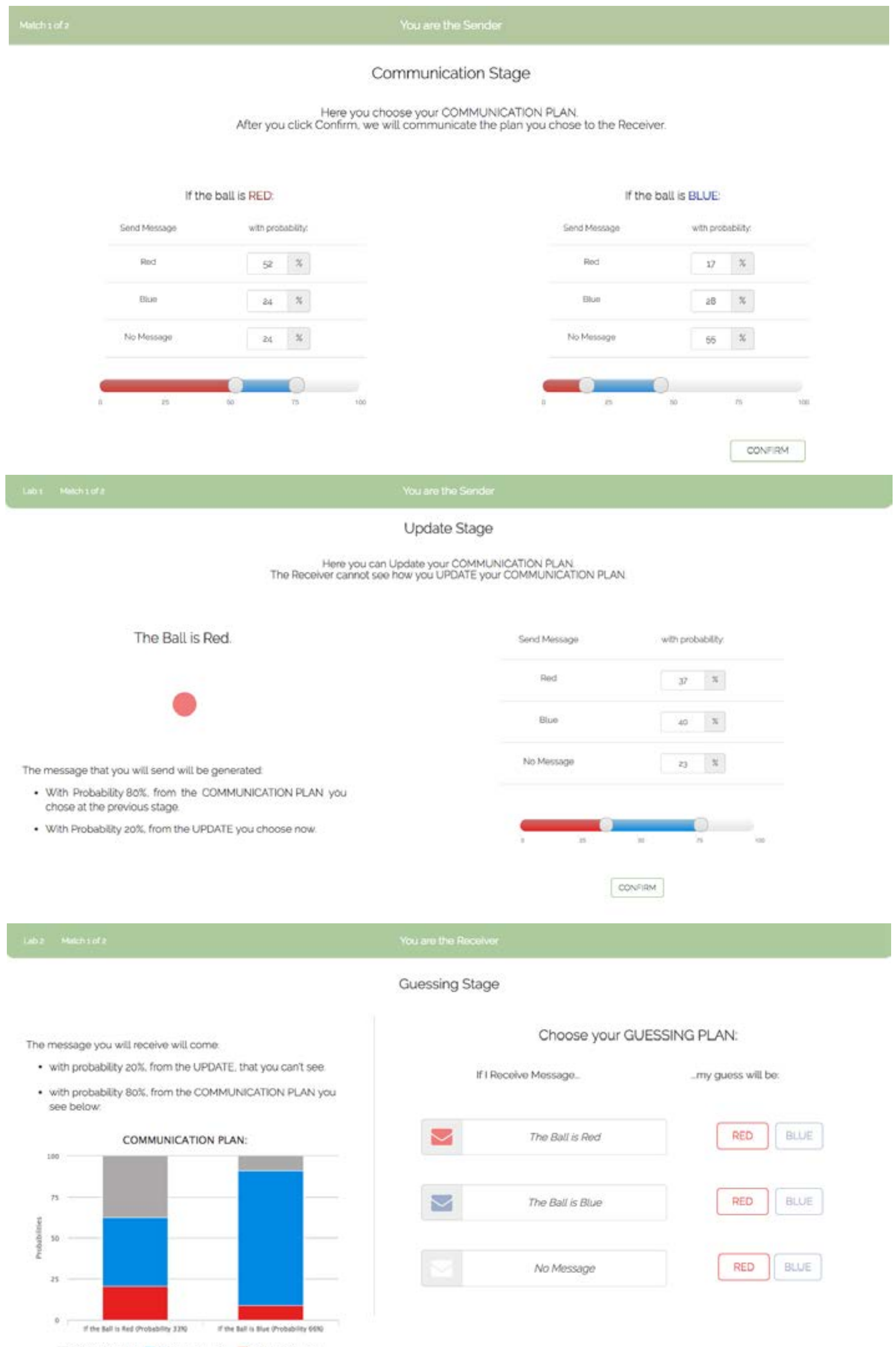


Figure E22: Sample Screenshots, $U80$: Commitment, Revision, and Guessing Stages

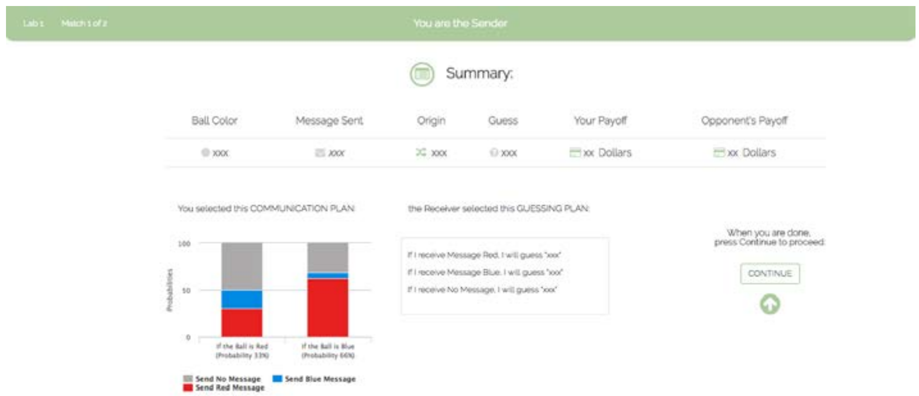


Figure E23: Sample Screenshots, $U80$. Feedback

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.

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 E10.

	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 E10: Payoffs

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 E11.

If Ball is Red:

- Message: "*The Ball is Red.*"
- No Message.

If Ball is Blue:

- Message: "*The Ball is Blue.*"
- No Message.

Table E11: Messages

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.
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.

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.

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.

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.

How is a message generated?

See attached table.

Practice Rounds:

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)

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.

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.