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 **1**

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

B.2 $U100S$ – Simplifying the Message Space 2

C A Closer Look at Receivers' Behavior **4**

C.1 Threshold Strategies in Main Treatments 7

D Additional Material **8**

D.1 Remaining Proofs 8

D.2 Correlation and Blackwell Informativeness 11

D.3 Examples that Fail the Refinement 14

D.4 Statistical Tests 16

D.5 $V0$ and $U0$ 17

D.6 Receivers' Behavior and Revealed Information 18

D.7 Gaussian Mixture Model 19

E Design **20**

E.1 Graphical Interface 20

E.2 Sample Instructions 20

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.

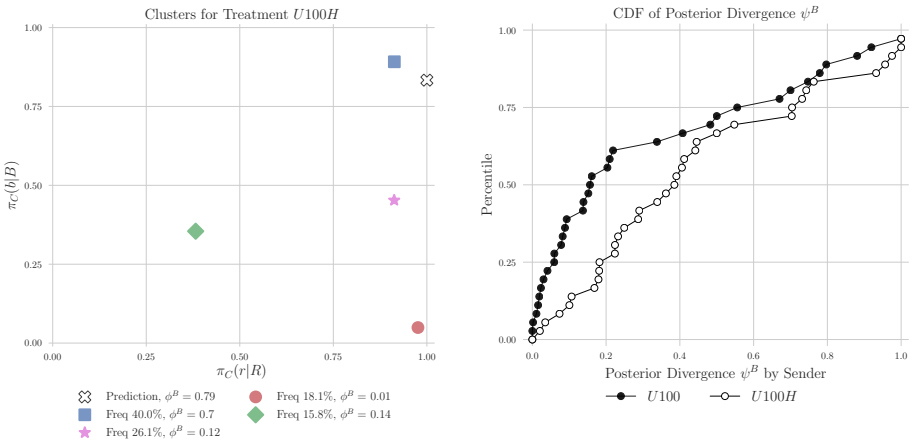

Figure B9: Strategy Clusters (left) and CDFs of Posterior Divergence ψ^B (right)

(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

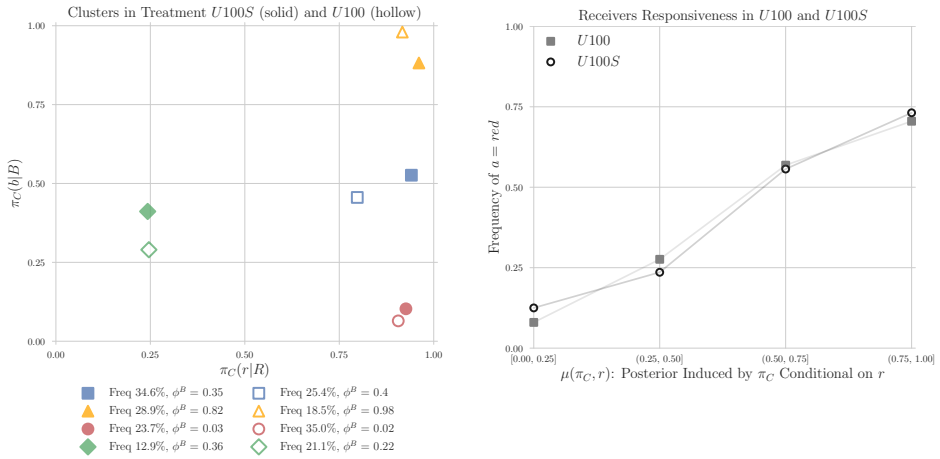

Figure B10: Senders' (left panel) and Receivers' (right panel) Behavior in $U100$ and $U100S$

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).

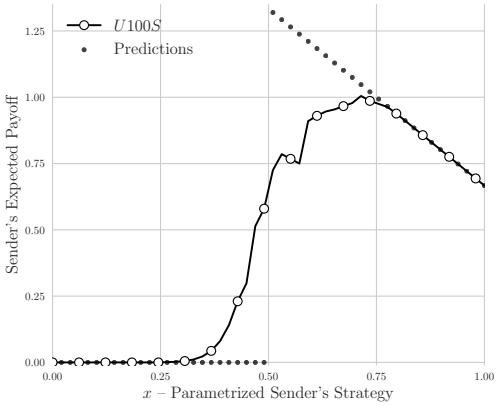

Figure C11: Probability of Guessing Red by Posterior and Message

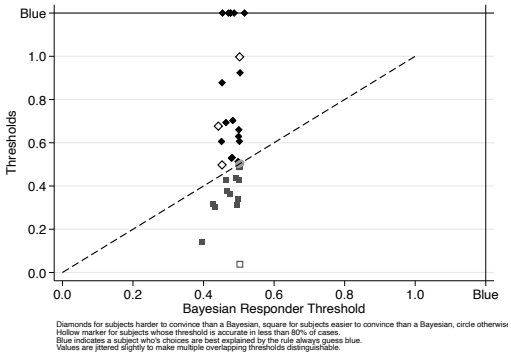

Figure C12: Estimated Thresholds: Actual Receivers vs Bayesians

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).

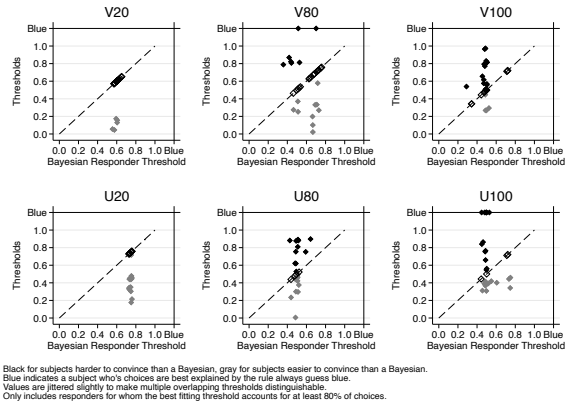

Figure C13: Estimated Threshold: Actual Receivers Against Bayesian

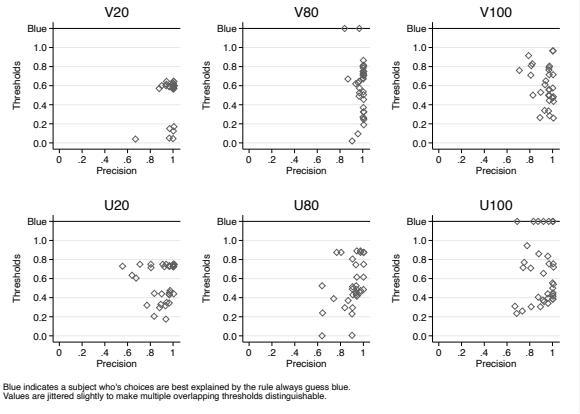

Figure C14: Estimated Threshold and Precision

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 (relatedly, see also (Ambuehl and Li, 2018)). 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.

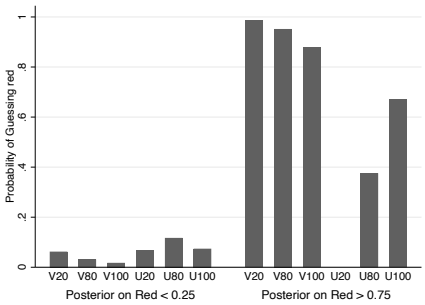

Figure C15: Frequency of $a = red$ for All Messages Given Posterior

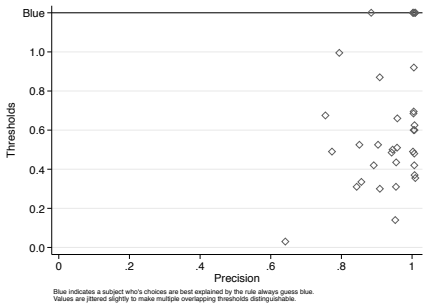

Figure C16: Estimated Threshold and Precision for Treatment U100S

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 $\varnothing \neq M' \subsetneq M$ 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. □

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 $\mu_0 = \frac{1}{3}$ and $q = \frac{1}{2}$. Let π be uninformative, in the sense that $\mu(m) = \mu_0$ for all m. Let π' induce posterior $\mu(m) = \frac{2}{5}$ with probability $\frac{5}{6}$ 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.

D.3 Examples that Fail the Refinement

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. 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 D5. 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 D5

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 D5, 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 D5 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 D6. 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 40). 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 $\pi_R(\theta_H|\theta_H) = 1 \neq 0$

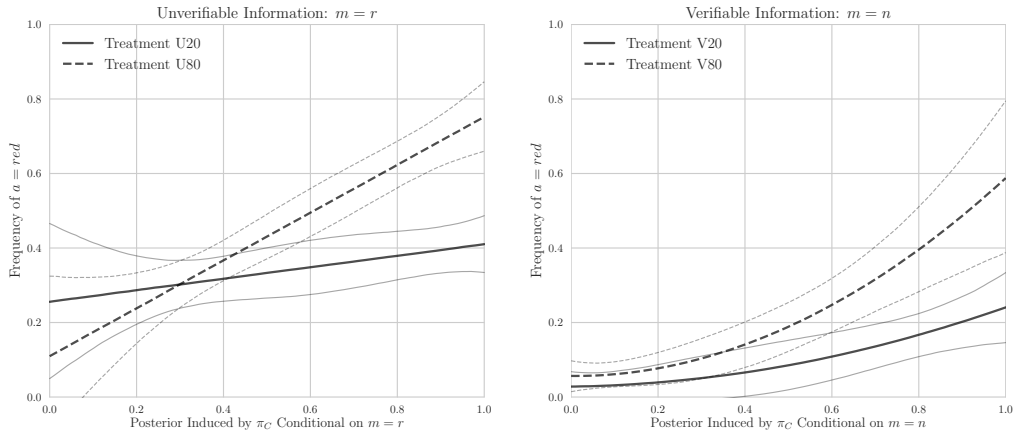

Figure D17: Receiver's Response to Persuasive Messages: $\rho = 0.2$ vs. $\rho = 0.80$

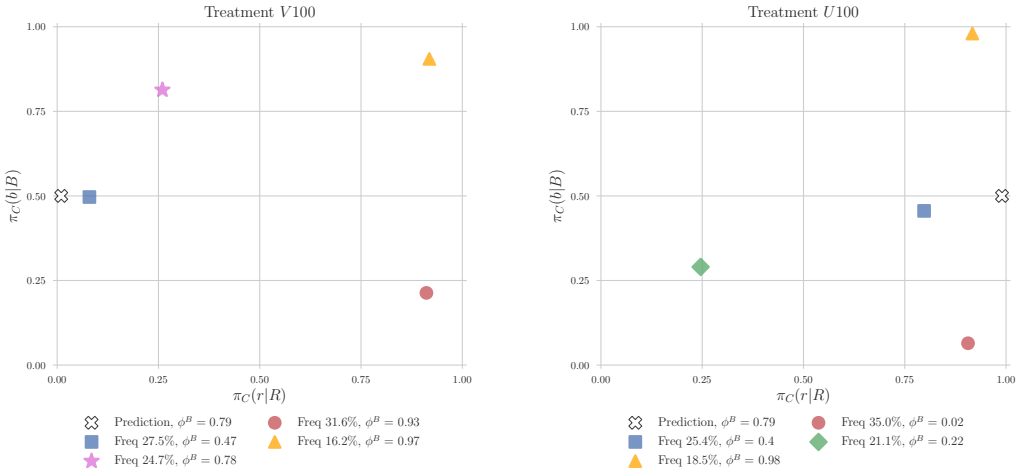

Figure D18: *k*-Means – Representative Strategies in Treatments with Full Commitment

Table D6

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

D.4 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).

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

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

Table D7: p-Values of Statistical Tests

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

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.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 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 $a, a' \in$

[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

$\{Red, Blue\}$, 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 D19 plots the posterior assignments of the clusters computed in that fashion for treat-

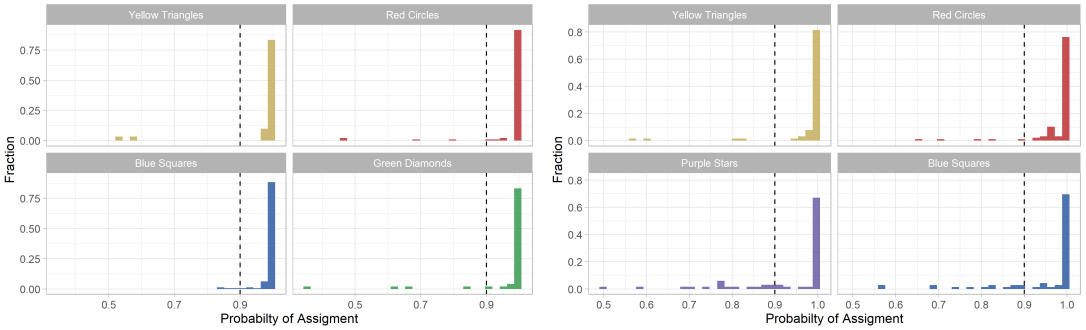

Figure D19: Posteriors Probabilities of k-Means Assignments for $U80$ (left panel) and $V80$ (right panel)

ments $U80$ (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 E20 and E21 shows the software interface of our experiment. More specifically, Figures E20 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 E21 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 E20 and E21 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 decisions of others, and partly on chance. On top of what you will earn during the session, you will receive an

Figure E20: Sample Screenshots, $U80$: Commitment, Revision, and Guessing Stages

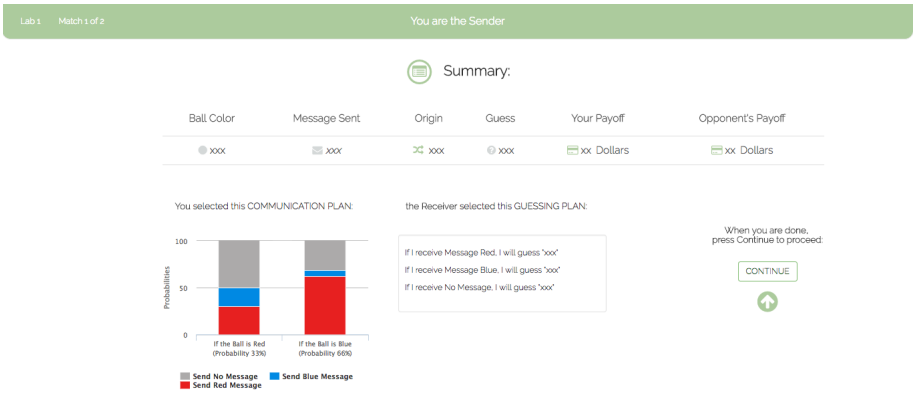

Figure E21: Sample Screenshots, *U*80. Feedback

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

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

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

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.