A Simple Experiment on Simple Bayesian Persuasion*

Pak Hung Au† OSub Kwon‡ King King Li§

February 1, 2023

Abstract

This experiment tests Bayesian persuasion (Kamenica and Gentzkow, 2011) in a simple setting. We adopt an experimental design in which the Sender chooses a partition of the state space. We find that 1) the Senders' strategies generally satisfy the optimal property that the weaker signal is fully revealing, but 2) their strategies are persistently suboptimal in the sense that the stronger signal is systematically too weak, resulting in a high rate of persuasion failure. However, 3) once we replace the Receivers with a robot who plays a known strategy, most Senders quickly learn to play the optimal strategy. This suggests that the key strategic element of Bayesian Persuasion is easy to understand for the Senders, although determining the posterior probability needed to persuade a human Receiver to take the desired action is a more difficult problem. We discuss some sources of the difficulty and provide evidence for them.

1 Introduction

In both political and economic realms, there are many settings in which one party attempts to persuade another party into taking an action by providing information. Among these is information design, or Bayesian Persuasion (Kamenica and Gentzkow, 2011). In this setting, the Sender chooses the information structure available to the Receiver, and the Receiver sees the realization without any distortion.

* This paper is a combination of two of our previous works, one titled "Bayesian Persuasion and Reciprocity: Theory and Experiment" by Pak Hung Au and King King Li, and the other one titled "Bayesian Persuasion in the Lab" by OSub Kwon. We thank Yaron Azrieli, Paul Healy, John Kagel, Dan Levin, James Peck, John Rehbeck, Renkun Yang, Huanxing Yang, and Xu Zhang for their invaluable advice. All errors are our own.

†Department of Economics, Hong Kong University of Science and Technology. Email: aupakhung@ust.hk.

‡School of Finance, Nankai University. Email: kwonosub@nankai.edu.cn.

§ Shenzhen Audencia Financial Technology Institute, WeBank Institute of Financial Technology, Shenzhen University. Email: likingking@gmail.com.

As a normative theory, the theory of information design tells us how the Sender can choose a particular information structure best achieve a certain goal. However, in practice, choosing an information structure means that the Sender needs to choose from an enormous set of complicated objects (signal spaces and conditional probability functions). Moreover, the decision can be further complicated by many other sources of difficulties in real life[1]. Is it possible for a real person to solve such a problem? Answering this question is important in understanding to what extent the theory of Bayesian persuasion is relevant to real-world phenomena as a descriptive theory.

In this experimental study, we test Bayesian Persuasion in a minimal setting to investigate, absent all other sources of difficulties, to what extent the key strategic element of the theory is alive. To minimize all other sources of difficulties that are not essential to the key strategic element, we design an experiment in which each Sender chooses a partition of the augmented state space, which Green and Stokey (1978) show is an equivalent way to represent an information structure to a state-contingent signal distribution. The key feature of this design is that neither Sender nor Receiver needs to use the Bayes formula to update their belief implied by a signal realization. This isolates the key strategic element of Bayesian Persuasion and massively reduces the complexity of the problem without distorting the meaning of Bayesian Persuasion or restricting the Sender's choice of information.

To understand how choosing an information partition works, consider a college (Sender) that wants to induce an employer (Receiver) to hire one of its graduates (as in, e.g., Boleslavsky and Cotton (2015)). This graduate comes from a pool of 40 good students and 60 mediocre students in the college. The employer is willing to hire a graduate if and only if the graduate has a grade that indicates that she is more likely to be a good student than a mediocre student. To maximize student placement, the college designs a grading policy that essentially partitions the student pool into different subgroups characterized by grades (e.g., those who get Grade A and those who get Grade B).

As it turns out, the college can get the employer to hire more students than the actual number of good students. The trick is twofold. First, by categorizing many good students as Grade A, the college can make Grade A a stronger signal (that the student is a good one) than Grade B. Because the weaker signal (Grade B), which does not induce employment, has no benefit to the college, it must give all the good students Grade A to ensure that all the good students are employed. This makes Grade B an unambiguous signal that the student is only a mediocre one. Second, because the strength of the stronger signal (Grade A) has no use beyond inducing employment, and because the objective is to maximize student

[1]For example, both the number of states and the signals can be extremely large, the Sender may not know how each signal is interpreted by the Receiver, and the Sender may not know the Receiver's preference, etc..

placement, the college must also pool some mediocre students into Grade A. This increases the frequency of Grade A and depresses its strength as a signal, but does not hurt the college as long as it still induces employment. Thus, in this example, the optimal grading policy for the college is to give Grade A to all 40 good students and 39 mediocre students (and Grade B to 21 mediocre students). This way, Grade A induces a posterior slightly higher than 50% (40/79) and the college can get the employer to hire 79% of its students.

We bring this simple case into the lab to understand how subjects perceive the basic strategic element of Bayesian persuasion. The optimal strategy of the Sender in this case always entails two features: 1) he should set the weaker signal to be fully revealing and 2) he should set the stronger signal just above the cutoff at which the Receiver is willing to take the Sender's preferred action. This logic applies regardless of where this cutoff is.

We find that the Senders' strategies generally satisfy the optimal property that the weaker signal is fully revealing. However, their strategies are suboptimal in the sense that the stronger signal is systematically weaker than what the Receivers require to take the Senders' preferred action, resulting in a persistently high rate of persuasion failure. This comes as a surprise because setting too low a posterior is a much more costly mistake than setting too high a posterior. This suggests the Senders may have difficulty in figuring out the posterior demanded by the Receivers for taking their preferred action.

The strategic uncertainty concerning the Receivers' decision rules appears to be the only factor that hinders the Senders from persuading optimally. When the role of Receiver is played by a robot with a known cutoff strategy, most Senders are able to choose the optimal strategy. This suggests that Bayesian Persuasion, a supposedly difficult problem, is easy to understand once all other sources of difficulties that are not essential to the key strategic element are lifted, although correctly determining what posterior a human Receiver would demand to take the preferred action remains a challenging problem to human Senders.

We also find that Receivers tend to demand more information to take the Sender's preferred action when the information structure is designed by human Senders than when it is randomly drawn by the computer. This suggests that social preference plays some role in the Receiver's decision that makes it even more difficult for the Senders to determine the Receivers' decision rule and therefore persuade them optimally.

There are also several other lab experiments studying some applications of bayesian persuasion. For example, Aristidou, Coricelli, and Vostroknutov (2019) compare the effectiveness of information design and mechanism design. They find that the Sender is more successful in inducing the Receiver to take his preferred action if he provides informational rather than monetary incentive. Wu and Ye (2022) study competitive persuasion and find that the competition between two Senders can increase information received by the Receiver. Meng and

Wang (2022) study how information avoidance behavior is affected by whether the information source is exogenous or endogenous (i.e., chosen by a Sender) and find that information avoidance is more prevalent than theoretical prediction.

We believe that Bayesian Persuasion is understudied in the lab partly because it is difficult to implement in a way that is easily interpretable to inexperienced human subjects. Consequently, we are unsure whether we can obtain meaningful results from a lab experiment. Our main contribution to the literature is to show that, by using an experimental design that exploits the well-understood partition interpretation of information, Bayesian Persuasion can be massively simplified to a point at which most people can understand. Consequently, our simple experimental design allows us to get very sharp findings and help us identify several important behavioral patterns that are not identified in previous experiments.

As a laboratory test of a communication model, our experiment is related to a large experimental literature on cheap talk[2] and (verifiable) information disclosure.[3] What distinguishes our experiment from these experiments is that the Receiver should take the "message" from the Sender at face value because the Sender can neither distort (as in cheap talk) nor hide (as in information disclosure) the realization of the signal, even if this is against the latter's interest.

In our experiment, the Sender essentially chooses a menu of lotteries for the Receiver. To do this optimally, the Sender needs to correctly predict the Receiver's risk attitude. Thus, our experiment is related to studies that examine how one predicts other people's choices under risk[4]. Some studies in this literature find that people predict that others are more risk-seeking than they actually are. These studies are consistent with our findings that the Senders set the stronger posterior as if they think the Receivers are more risk-seeking than they actually are. However, because the evidence is mixed in this line of literature (with the exception of the common finding that people are not good at predicting other people's risk attitudes), and because our setting is different, it is not clear how the findings of these studies can be applied here.

The rest of the paper is organized as follows. In Section 2, we outline the basic structure of the experimental design, its theoretical interpretation, and our theoretical predictions. In Section 3, we report our main results. In Section 4, we briefly discuss the results and how the design can be extended.

[2]See, for example, Dickhaut et al. (1995), Forsythe et al. (1999); see Blume et al. (2017) for a survey on this topic.

[3]See, for example, Jin et al. (2017), Hagenbach (2018).

[4]See, for example, Hsee and Weber (1997, 1999) , Siegrist et al. (2002), Eckel and Grossman (2008)

2 Experimental Design

2.1 Part 1: Persuasion Game

The basic structure of our persuasion game is summarized as follows. In each round of the game, a Sender (Role 1) is randomly matched with a Receiver (Role 2). The roles are fixed throughout the experiment. The task that the two parties face concerns the color of a ball randomly drawn. The Sender's goal is to induce the Receiver to guess Red, whereas the Receiver's goal is to guess the color of the drawn ball correctly.

The persuasion game consists of four steps.

Step 1: A ball is randomly drawn among 40 red balls and 60 blue balls. Neither the Sender nor the Receiver knows the color of this drawn ball until the end of a round.

Step 2: The Sender chooses how to divide the 100 balls into two boxes, Box A and Box B[5]. Note that, at this point, exactly one box must contain the drawn ball in Step 1.

Step 3: After the Sender's decision, Box A or Box B, whichever contains the drawn ball, is assigned to the Receiver. The composition of the assigned box (the numbers of red and blue balls) is revealed to the Receiver.

Step 4: The Receiver submits a guess about the color of the drawn ball. Note strategy method is used in this step to elicit the Receiver's response to each box regardless of the realization of boxes.

Some justification for the design is worth noting. In the Kamenica and Gentzkow (2011) setting, the Sender chooses a state-contingent signal distribution, which is defined as a mapping from the payoff relevant state space to the set of signal distributions. Here in our persuasion game, the Sender chooses a partition of the state space (i.e., the 100 balls) that has both the payoff relevant dimension (i.e., the color of the drawn ball) and the payoff irrelevant dimension[6] (i.e., the identity of the drawn ball). Green and Stokey (1978) show that these are two equivalent representations of information structure and in fact, both notions are used in Bayesian Persuasion literature (e.g., Brooks et al. (2019)). Thus, the Sender in our persuasion game indeed faces a Bayesian persuasion problem.

This partition representation of information directly reflects Bayesian updating. To see this, note that in the beginning, each of the 100 balls is equally likely to be the drawn ball. Suppose after the Sender chooses a partition, the Receiver is assigned Box A so that she knows that Box A contains the drawn ball for sure. Although this information rules out the

[5]In practice, he chooses how many red balls and blue balls to be put into Box A. Then, all the rest of the balls are automatically put into Box B.

[6]The payoff irrelevant dimension is needed and should be rich enough to allow the Sender to freely choose how states and signals are correlated.

balls in Box B being the drawn ball, it does not change the fact that each ball in Box A is still equally likely to be the drawn ball, which is exactly what is required by Bayesian updating. The fact that our experiment directly reflects the process of Bayesian updating makes the Sender's action highly interpretable and, more importantly, enables the Sender to choose the distribution of posteriors directly. Specifically, the probability that a box contains the drawn ball is simply the number of balls put in this box divided by 100, and the posterior belief that the drawn ball is red given a box is simply the fraction of red balls in this box. The Bayes plausibility constraint that the expected posterior probability must be equal to the prior probability is automatically satisfied.

Payoff: The Sender gets a payoff of HK$70 if the Receiver guesses Red and a payoff of HK$0 otherwise. The Receiver's payoff depends on a treatment variable $L \geq 1$, and it is summarized as follows[7]:

$$\begin{cases} \text{HK\$40} & \text{if the guess is correct and the color is Red} \\ \text{HK\$40}L & \text{if the guess is correct and the color is Blue} \\ \text{HK\$0} & \text{otherwise} \end{cases}$$

That is, the Receiver's payoff is HK$0 whenever the guess is incorrect. But when the Receiver guesses correctly, the prize of a correct Red guess can be different from the prize of a correct Blue guess. The higher the L, the larger the prize of a correct Blue guess, and therefore the more inclined the Receiver is to guess Blue at each posterior belief.

Feedback: At the end of a round, each subject is told how the Sender divides the balls, how the Receiver guesses under each box, the actual color of the drawn ball, the box that is actually assigned, and his or her own payoff.

2.2 Part 2: Individual Task

After the persuasion game, each subject faces an individual task depending on his or her role in the persuasion game. In this individual task, if the subject is a Sender, he guesses what the Receivers have guessed in a previous (randomly drawn) round of the persuasion game, and his goal is to match the guess of the Receiver's guess in that round. The purpose of this task is to understand the Sender's belief about the Receiver's response. If the subject is a Receiver, she again guesses the color of the drawn ball in a previous (randomly drawn)

[7] A popular payoff scheme used by authors such as Frechette et al. (2022) equalizes the rewards for correct Red and Blue guesses but varies the punishment for incorrect Red and Blue guesses. We choose our payoff scheme over that popular one because we believe ours makes the Receiver's problem look simpler from the Sender's perspective.

round of the persuasion game, and her goal is still to match the color of the drawn ball. The purpose of this task is to understand what the Receiver would have guessed if her action had no influence on the Sender's payoff.

The individual task consists of three steps:

Step 1: For each subject, we construct the set of rounds of persuasion game in the same session in which this subject did not participate, that is, the set of rounds played only by other subjects in the same session.

Step 2: From this set, we randomly pick a round and replicate its Step 1-3: we randomly draw a ball from the 100 balls, divide the 100 balls into two boxes as the Sender of that round has done, and assign the box that contains the drawn ball to the subject. Similar to the persuasion game, the strategy method is used in this step to elicit the subject's response to each box regardless of the realization of boxes.

Step 3: If the subject has played the persuasion game as a Sender, he needs to guess the choice of the Receiver in that round when presented with the box. If the subject has played the persuasion game as a Receiver, she needs to guess the color of the drawn ball given that the drawn ball is in that box (the same task as in the persuasion game).

Payoff: The Sender gets a payoff of HK$40 if his guess about the Receiver's choice is correct and a payoff of HK$0 otherwise. The Receiver's payoff structure is the same as the persuasion game and also depends on the treatment variable L. Unlike the persuasion game, neither the Sender nor the Receiver has any influence on another subject's payoff in this part. Note that because the rounds faced by subjects in this part are those they have not participated in, they could not use their feedback information in the persuasion game to improve their decisions in this part.

Feedback: At the end of a round, the Receiver gets the same feedback as in the persuasion game, but the Sender is not provided any feedback, which may influence his belief about the Receivers' behaviors.

2.2.1 Theoretical Predictions

After being shown the composition of the drawn box, the Receiver decides between guessing red or blue, each of which represents a lottery. It is immediate, though noteworthy, that the only payoff-relevant factors for her are the lotteries themselves. Neither the identity of the Sender, nor Whether her decision affects the Senders' payoff, are payoff-irrelevant to the Receiver. Conditional on the composition of the drawn box, a perfectly-rational Receiver would therefore behave identically in both the individual tasks and the persuasion game.

Given the Receiver's payoff structure, it is easy to verify that L is the threshold likelihood

ratio at which a risk-neutral Receiver is indifferent between guessing Red and guessing Blue. More precisely, the Receiver collects a higher expected payoff by guessing Red if and only if $Pr(Red)/Pr(Blue) \geq L$ (or, equivalently, $Pr(Red) \geq L/(1+L)$ in terms of the posterior probability.)

Importantly, when $L = 1$, guessing the more likely color stochastically dominates guessing the other color, and therefore the cutoff posterior is 50% regardless of risk attitude. However, when $L > 1$, the prize from a correct Blue guess is larger than the prize from a correct Red guess, thus raising the cutoff posterior above 50%. As a risk-averse Receiver may choose a small prize with a high probability over a large prize with a small probability, her cutoff posterior for choosing Red is lower than that of a risk-neutral Receiver. In contrast, the cutoff posterior of a risk-loving Receiver is higher than a risk-neutral Receiver.

Prediction 1: The Receiver should base her guess only on the proportion of red balls in the drawn box. Moreover, she follows a cutoff strategy: guessing red if and only if the posterior (that the drawn ball is red) of the box is no less than some threshold. Moreover, this cutoff is increasing in L.

For the Sender, the problem is more complicated, so for ease of exposition, we introduce some definitions. As the overall fraction of the red balls is equal to the prior probability, if one box has a fraction higher than the prior, the other must have a fraction lower than the prior. Throughout, we will call the former box *the stronger box* and the latter box *the weaker box*. We will call the corresponding posterior belief *the stronger posterior* and the *the weaker posterior*, respectively.

As the weaker posterior is below the prior, it has no hope of persuading a rational Receiver to guess red. The only hope for successful persuasion lies in the stronger box. As such, any red ball left in the weaker box is essentially "wasted." Transferring the red ball to the stronger box raises the strength of the stronger box, as well as its frequency of being drawn. This results in a (weak) improvement in the probability of successfully persuading the Receiver to guess red.

Prediction 2: The Sender always puts all the red balls in the stronger box. That is, the weaker posterior should be equal to 0.

While the Sender has the option of putting only red balls in the stronger box, thus making the stronger posterior 1, he can improve his chance of successful persuasion by putting some blue balls in the stronger box, as long as the stronger posterior is kept above the Receiver's cutoff. This works by raising the likelihood of the stronger box being drawn, and hence a red guess is made by the Receiver. In fact, if the Receiver's cutoff strategy is known to the Sender, the latter's optimal strategy that maximizes the likelihood of persuasion has the stronger posterior just enough to induce a red guess. Even if the Receiver's cutoff is

not known to the Sender, the Sender can secure a red guess following the stronger box by completely separating the red balls and the blue balls, thus inducing an (ex-ante) probability of red guess that coincides with the prior probability.

Prediction 3: The Sender should put some blue balls in the stronger box in addition to all the red balls so that the stronger box is just enough to induce the red guess. Consequently, the stronger box always induces a red guess by the Receiver. Moreover, the (ex-ante) probability of the red guess is no lower than the prior probability.

A crucial message of Prediction 3 is that in equilibrium, the stronger posterior offered by the Sender always succeeds in persuasion. In setting the stronger posterior, the Sender faces a discrete jump in his payoff around the Receiver's cutoff: when the stronger posterior is slightly above the cutoff, the Sender induces the red guess with the highest probability; but when the stronger posterior is slightly below the cutoff, the Sender induces the red guess with probability 0. Thus, it is usually more costly for the Sender to set the stronger posterior too low than to set the stronger posterior too high.

2.3 Treatments and Procedure

Table 1 summarizes the treatments of the experiment. In each treatment, the subjects face two values of L in a session, 10 rounds for each value L in each part. That is, for the treatment with both the persuasion game and the individual task (Treatment 1), each session has 20 rounds of persuasion game and 20 rounds of individual task, with 10 rounds for each value L. For the treatments with the persuasion game only (Treatment 2 and 3), each session has 20 rounds of persuasion game, also with 10 rounds for each value of L.

Table 1: Treatments

	Receiver	L	Theoretical Prediction	Sessions (No. of Subjects)
Treatment 1	Human	1; 3	(40, 39); (40, 13)	4 (16, 20, 20, 16)
		3; 1	(40, 13); (40, 39)	4 (20, 12, 20, 16)
Treatment 2	Robot	1; 3	(40, 39); (40, 13)	1 (14)
		3; 1	(40, 13); (40, 39)	1 (17)
Treatment 3	Random Robot	1; 3	(40, 26); (40, 7)	2 (18, 14)

We recognize that the case of L = 1 is an interesting but knife-edge case where the best response of the Receiver is independent of her risk attitude. Thus, we also include the case of $L = 3$ to see whether there is any qualitative difference compared to the $L = 1$ case. Adopting a within-subjects design allows us to determine whether the Sender has a robust understanding of the problem and how responsive Senders are to the change in the incentive faced by the Receivers.

In Treatment 2, we replace the human Receivers with a robot Receiver. Unlike a human Receiver, a robot Receiver plays a cutoff strategy known to the Sender: the robot guesses Red if and only if the fraction of the red balls in the urn it receives is higher than a publicly announced x, where x solves $L = x/(1-x)$. This robot treatment serves as an important benchmark for testing, absent strategic uncertainty from the Receiver side, whether the Senders can design information optimally. Note that in this treatment and the next, we do not include the Sender's individual task as there is no meaningful belief about the Receiver to elicit.

In Treatment 3, we introduce a different robot Receiver that still guesses Red if and only if the fraction of the red balls is higher than a certain cutoff, but the cutoff is not revealed to the subjects. Specifically, this cutoff is drawn from the uniform distribution $U(x, x+0.1)$, which is publicly announced. This treatment (which has random cutoffs drawn from a known distribution) can be viewed as a middle case between Treatment 1 (which has cutoffs determined by human Receivers) and Treatment 2 (which has deterministic cutoffs). It is straightforward to verify that in this treatment, for each x, it is always the best response to set the stronger posterior at $x+0.1$, which reflects the fact that it is generally more costly for the Sender to set the stronger posterior too low than too high[8]. Note that in this treatment there is no obvious anchor that provides any hint to the Sender about the optimum strategy.

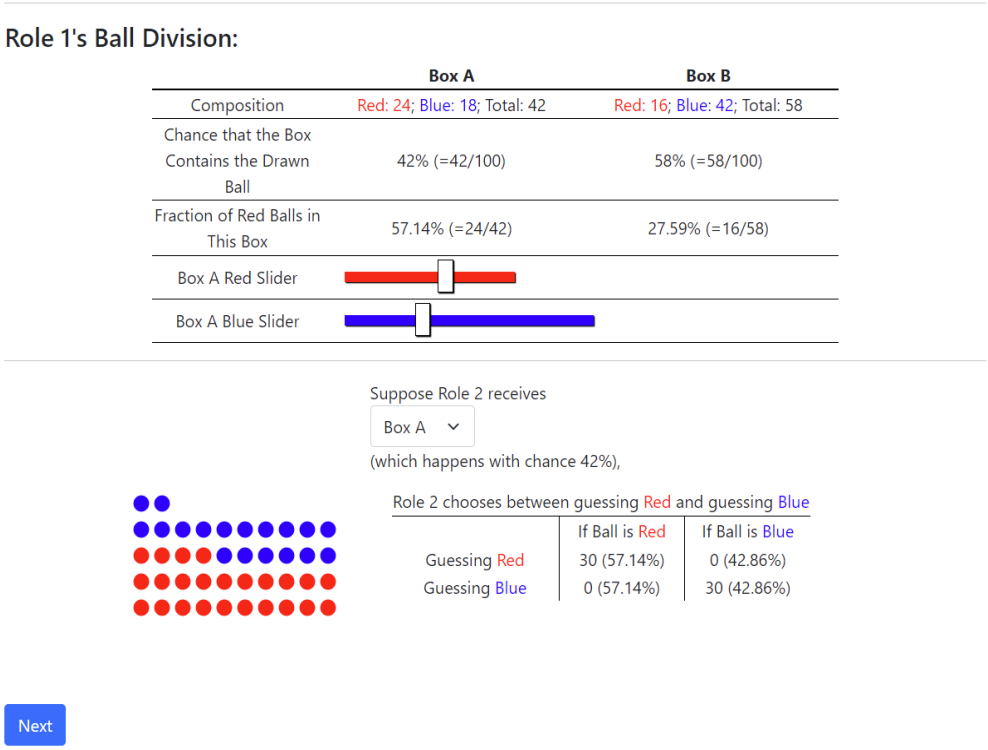
It is worth remarking that in both Treatment 2 (Robot) and Treatment 3 (Random Robot), the Sender is essentially facing a single-agent decision-making problem. Moreover, given the robot behaviors we specify, the optimal choice of the Sender satisfies the prescription in Prediction 2, 3, and 4 in the previous section. Comparison across these treatments allows us to study the effect of strategic uncertainty (arising from the Receiver's choices) on Senders' behaviors.

We ran our experiment online in December 2021 and March 2022. Subjects were recruited through emails and were predominantly undergraduate students at HKUST. Each session included 12-20 subjects. At the beginning of a session, the experimenter read the instructions aloud to the subjects. Then, the subjects played 2 practice rounds, with 1 round for each L. In the Human Receiver treatments, the subjects played both the role of Sender and Receiver in these practice rounds. That is, they chose a strategy as a Sender and reacted to it as the Receiver. The purpose of this design is to let all the subjects gain a better understanding of the problem each role faces, while preventing them from learning how to play the game from others. In the two Robot Receiver treatments, the subjects simply played as the Sender for two rounds. After the practice rounds, the subjects played 10 rounds for each L, with the

[8]To show this, simply note that the probability of Red guess $\frac{0.4}{p}\frac{p-x}{0.1}$ is increasing in p within $[x, x+0.1]$

role of each subject being fixed throughout the session. Each session of the Human Receiver treatment lasted approximately 90 minutes and each session of the two Robot Receiver treatments lasted approximately 75 minutes. For every 10 rounds of each part, we randomly picked one round for the payment in addition to a HK$75 participation fee. This resulted in an average payment of about HK$225 for each subject in the Human Receiver treatments and HK$148 in the two Robot Receiver treatments.

Figure 1: Screenshot of Sender's Interface in Treatment 1



It is worth mentioning how the Sender's problem is exactly presented as we are interested in how he solves it. Figure 1 is a screenshot of the Sender's interface in Treatment 1. As the screenshot shows, the Sender chooses how many Red and Blue balls to put in Box A by moving two sliders. To reduce possible anchoring effects, each slider is programmed so that it does not appear on the slider bar until the Sender clicks on the slider bar. To make the relative number of Red and Blue balls put in Box A clear to the Sender, the length of each slider bar is set proportional to the total number of balls of the corresponding color. Whenever the Sender moves a slider, the interface automatically calculates how many balls of each color are put in Box B and all the relevant probabilities. The results of these calculations are shown to the Sender in real time. In the lower half of the Sender's interface, we show the Sender the problem faced by the Human Receiver exactly as is presented to the latter in

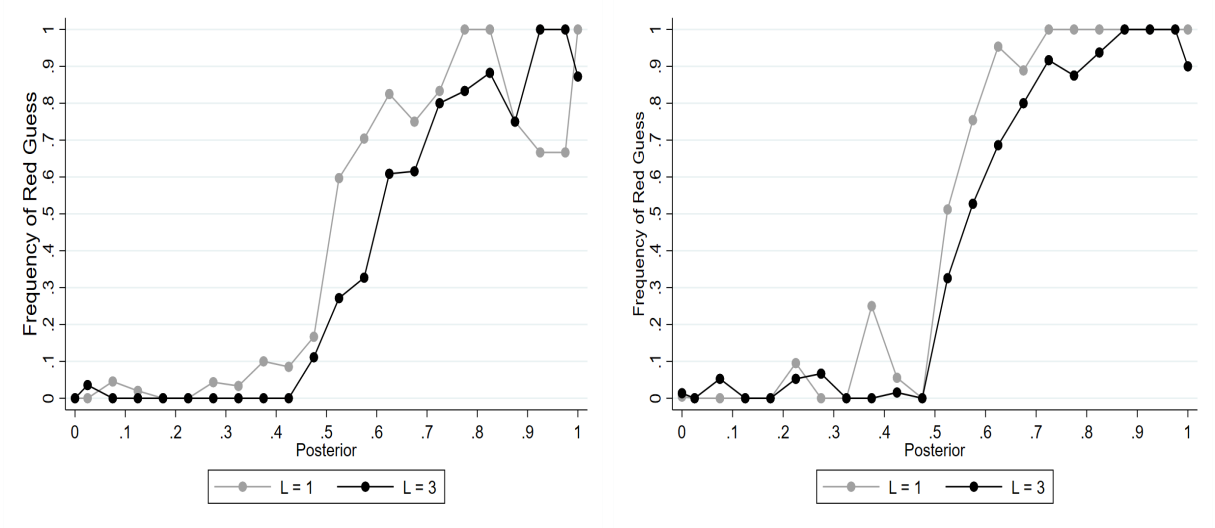
Treatment 1. In Treatment 2 and 3, this area displays the statement of the Robot Receiver's strategy. Appendix B has the full instructions of the experiment.

3 Results

3.1 Receiver's Behavior

We start with a discussion of the Receivers' behaviors in the Human Receiver treatments. Choosing between guessing red or blue is equivalent to deciding which corresponding lottery to pick. Establishing the regularity of the Receiver's behavior is important for us to evaluate the performance of the Senders in the Human Receiver treatments, and compare with that in the two Robot Receiver treatments.

Figure 2: Receiver Aggregate Response



(a) L = 1 then L = 3

(b) L = 3 then L = 1

Figure 2 depicts the aggregate response of the Receivers to different levels of posterior. It shows that, regardless of the order of L,[9] for each L, the aggregate probability of the red guess increases continuously in the posterior belief.[10] Furthermore, when the posterior is lower than 0.5, almost no Receiver guesses Red regardless of L, which is a reasonable response. Thus, the only interesting region of posteriors is $[0.5, 1]$.

[9]Our probit regressions show that there is only slight order effect when $L = 1$.

[10]In Figure 2, there is some non-monotonicity within the interval $[0.9, 1]$ We attribute this non-monotonicity to the lack of data because, as is noted in Section 3.4, the posteriors within this region are unpopular choices among the Senders.

Within this region, Receivers are generally responsive to the change in L. To show this, we partition this region into two subintervals ($[0.5, 0.75)$ and $[0.75, 1]$) and test whether the average frequency of red guesses in each subinterval differs across L. Table 2 shows the frequency of the Red guesses in each subinterval under each L. We see that there is a significant difference in the frequency of Red guesses between L's in $[0.5, 0.75)$ but less so in $[0.75, 1]$. This is reasonable because most Receivers guess Red for sufficiently high posteriors.

Table 2: Frequency of Red Guess in Each Subregion

	L = 1 then L = 3		L = 3 then L = 1	
	$[0.5, 0.75)$	$[0.75, 1]$	$[0.5, 0.75)$	$[0.75, 1]$
$L = 1$	67.56%	92.68%	72.52%	100%
$L = 3$	47.11% ($-20.45\%^{***}$)	87.36% (-5.33%)	58.60% ($-13.92\%^{**}$)	91.11% ($-8.89\%^{*}$)

Difference with the previous row reported in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$ (Two Sided t-Test, Clustered at the Receiver Level).

However, when $L = 1$, a significant portion of Receivers guess Blue under a posterior slightly higher than 50%, although doing so is stochastically dominated. This cannot be explained by random decision error (e.g. Quantal Response)[11] because the Receiver's payoff from guessing Red at a posterior slightly lower than 50% is symmetric to that from guessing Blue at a posterior slightly higher than 50%, and yet when the posterior is slightly lower than 50%, they almost always guess Blue. This asymmetric response of the Receivers around 50% is robust and is also found in our previous drafts (Au and Li, 2018, Kwon, 2020) with different subject pools. It is an indication that Receivers have some behavioral bias in our environment that is unrelated to their ability of Bayesian updating.[12] As the results in our next subsection show, there is some evidence that this phenomenon is related to subjects' social preferences. Another possible cause of this phenomenon is a Receiver's bias toward the prior probability (that the drawn ball is Red with a 40% probability) because the ball is drawn before Sender's ball division.

When $L = 3$, a risk-neutral Receiver would guess Red if and only if the posterior is higher than 75%, but from Figure 2, we see that more than half of the Receivers guess Red within the posterior region $[0.6, 0.7]$. Their behavior is consistent with risk aversion, as they choose a 70% chance of getting a small prize over a 30% chance of getting a large prize even though the former is lower in expectation.

[11] See Appendix A for full description of our Quantal Response Equilibrium (QRE) estimation.

[12] In Frechette et al. (2022), there is a treatment theoretically identical to our $L = 1$ treatment, but the Receivers are required to do Bayesian updating. In that treatment, in our terminologies, the Receiver's frequency of guessing Red increases slowly in the posterior probability and does not exhibit any asymmetry in response around 50%. We believe that the reason why their findings differ from ours is that the Receiver's failure of Bayesian updating is so severe that it makes it impossible to detect any other behavioral bias.

Result 1: Receivers respond monotonically to posterior and L. When $L = 1$, most of the Receivers guess Red when the posterior is higher than 50%, although their response is sharply asymmetric around 50% posterior. When $L = 3$, most of the Receivers guess Red when the posterior is higher than 60%, indicating risk aversion.

Table 3: No. of Receivers Compatible with Cutoff Strategies

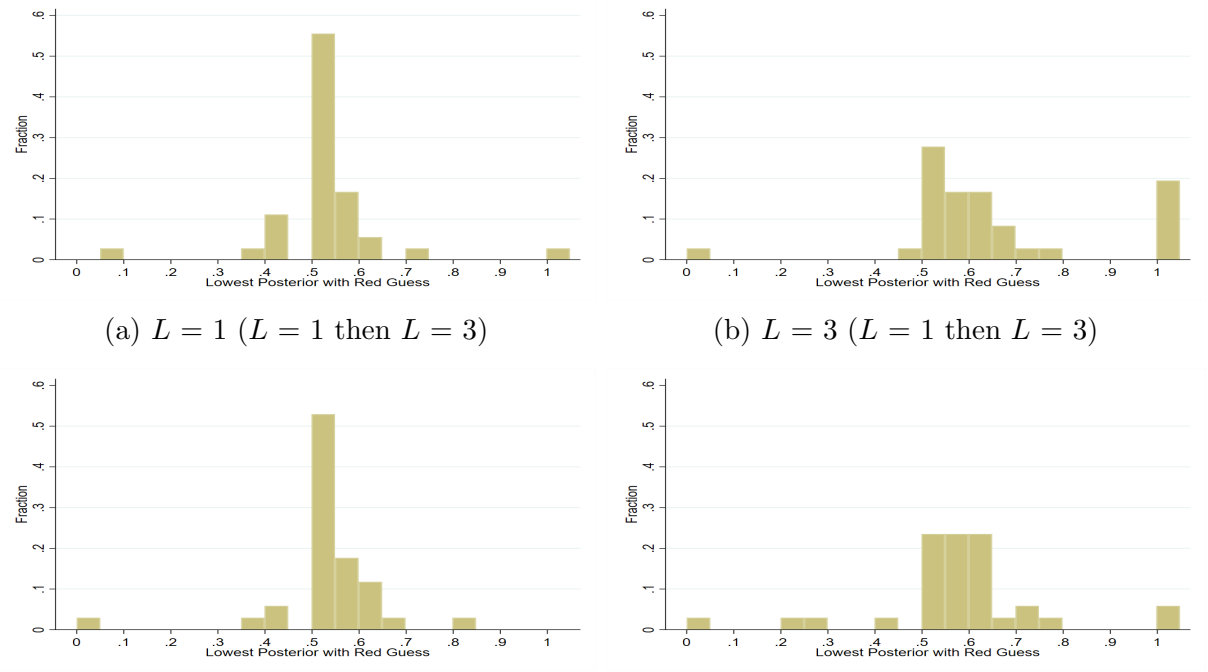
	L	0 Violation	≤ 2 Violations (Above)	≤ 2 Violations (Below)
L = 1 then L = 3	1	47.22%	75%	88.89%
	3	55.56%	86.11%	80.56%
L = 3 then L = 1	1	61.76%	88.24%	85.29%
	3	47.06	91.18%	82.35%

At the individual level, does a Receiver play a cutoff strategy? This question is important as the optimality of the Sender's strategy derived in Section 2.2.1 relies on the Receivers playing cutoff strategies. To answer this question, for each Receiver and L, we look for the highest posterior at which she guesses Blue,[13] and then we count how many times she guesses Red at a lower posterior. Similarly, we look for the lowest posterior at which she guesses Red and count how many times she guesses Blue at a higher posterior. This way, we obtain two measures of violations of the cutoff strategies that will coincide with each other if the Receiver strictly follows a cutoff strategy. Note that the two measures are very demanding. For example, if once by error a Receiver guesses Red under a weak posterior, we are likely to record many instances of her violation of the cutoff strategy. Table 3 shows that for each L, most Receivers have less than 2 violations both from above and below. In fact, more than 54.28% (51.43%) of the Receivers pass the even more demanding requirement to have no violation at all when $L = 1$ ($L = 3$). Thus, Receivers are generally consistent with a cutoff strategy.

Figure 3 shows the distribution of the lowest posterior with a Red guess. Although Receivers are mostly compatible with cutoff strategies, they do not use the same cutoff, which is the reason why the aggregate response is a continuously increasing function rather than a step function. In particular, when $L = 1$, most Receivers set their cutoffs around 0.5, but 31.42% of them still have estimated cutoffs higher than 0.55. This explains why the aggregate response is very steep around the region $[0.5, 0.55]$ but does not go all the way to 100% guessing red when the posterior is slightly higher than 0.5. Moreover, the average Receiver cutoff is higher at a higher value of L (the Wilcoxon ranksum test, $p < 0.01$), consistent with the theoretical prediction.

[13] If a Receiver empirically only guessed Blue, we take this posterior as 1.

Figure 3: Distribution of Lowest Posterior with Red Guess by L



(a) $L = 1$ ($L = 1$ then $L = 3$)

(b) $L = 3$ ($L = 1$ then $L = 3$)

(c) $L = 1$ ($L = 3$ then $L = 1$)

(d) $L = 3$ ($L = 3$ then $L = 1$)

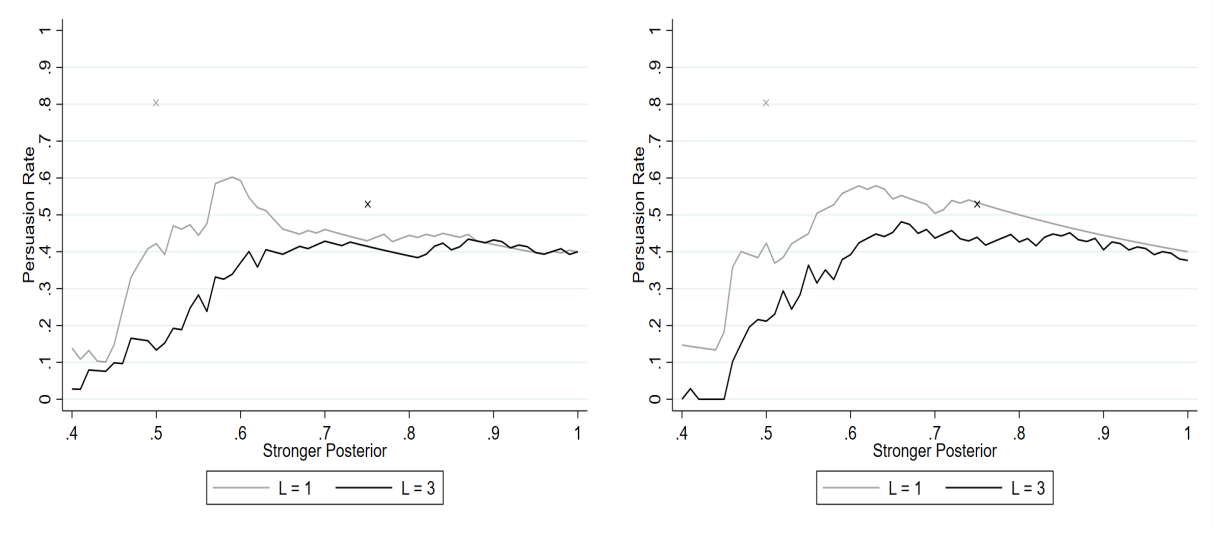
Result 2: Receivers generally follow cutoff strategies at individual level. Despite facing the same payoff function, there is significant heterogeneity in the cutoffs across Receivers. Moreover, the cutoffs are higher at $L = 3$ than $L = 1$.

The fact that in general, Receivers play cutoff strategies at the individual level means that, as is argued in Section 2.2, the optimal strategy of the Senders is to put in one box all the red balls plus some additional blue balls depending on the Receivers' responses. On the basis of these observations, we estimate the empirical best response of the Senders to the Receivers' strategies. For each L of the human treatment, we estimate the probability that the Sender can persuade a random Receiver in our experiment under the assumption that 1) he puts all the red balls in the stronger box and 2) he sets the posterior of the stronger box to be some p in $[0.4, 1]$. [14] Note that the number of red balls put in the stronger box and the stronger posterior together completely define a Sender's strategy.

Figure 4 shows the estimated persuasion rate as a function of the stronger posterior p, with $\times$ indicating the theoretical benchmark of a risk-neutral Receiver for each value of L. When $L = 1$, the estimated persuasion rate increases continuously from 0.4 to 0.6, beyond which it descends gently. When $L = 3$, it climbs continuously from 0.4 to 0.7, and remains

[14] More specifically, for each $p \in [0.4, 1]$ and for each Receiver we look at her action under her nearest observed posterior. We take the percentage of Red guesses under this p as our estimate.

Figure 4: Estimated Expected Persuasion Rate



(a) L = 1 then L = 3

(b) L = 3 then L = 1

more or less flat in the region above 0.7. In both cases, the empirical best response requires the Sender to set the stronger posterior at a point that can persuade most Receivers to guess Red. Equally importantly, we find that the estimated persuasion rate function is steeper on the left of the peak than the right. That is, under-informing Receivers is generally more costly than over-informing. This is reasonable because setting the stronger posterior lower than the Receiver's cutoff will never lead to persuasion. But setting the stronger posterior higher than her cutoff will only result in a lower frequency of the stronger posterior. This asymmetry of response leads to a kink in the expected persuasion rate function.

Result 3: Given the Receiver's strategy, the empirical best response generally requires the Sender to set the posterior high enough to persuade most Receivers.

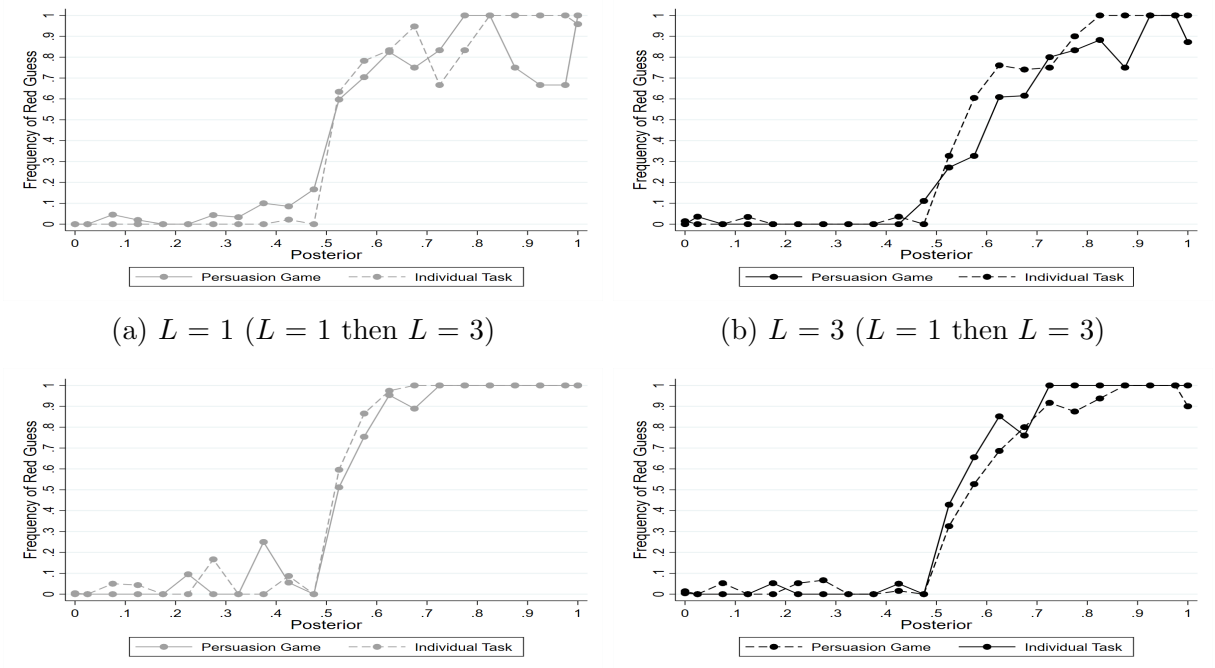
3.2 Receiver's Behavior in Game vs. Individual Decision

In this subsection, we compare the Receiver's response when they can influence the payoff of the Sender against that when they cannot. Recall that in our Human Receiver treatment, after the subjects finish playing the persuasion game, the Receivers are given an additional individual task in which they face an essentially identical lottery choice problem between guessing Red and guessing Blue as in the persuasion game, with one crucial difference. Specifically, in the individual task, the ball allocations of the boxes (i.e., the lottery problems facing the Receiver) are taken from the persuasion games that have already finished, and consequently, the Receiver's choices do not carry any externality. We use the Receivers'

choices in this individual task as a proxy of how Receivers would have acted were they cannot influence the Sender's payoff in the persuasion game. For effective comparison, we keep the framing of the decision problem as similar as possible to the persuasion game because the past literature has warned that the framing of the problem can heavily influence peoples' decisions under uncertainty.

If the Receiver's choices are guided only by his own material payoff, she should form her guess based solely on the posterior she faces. Her choice behaviors in the individual task should thus be identical to those in the persuasion game (condition on the proportion of red balls in the drawn box). However, the fact that the Receiver's action is able to influence the payoff of both the matched Sender and herself in the persuasion game, whereas this ability is absent in the individual task, suggests the possibility of reciprocation in the persuasion game. Specifically, in deciding whether to guess red and thus reward the Sender, a Receiver may take into account her perception of the matched Sender's kindness or hostility inferred from the Sender's offer of information.

Figure 5: Receiver Response Comparison (Persuasion Game vs. Individual Task) by L



(a) $L = 1$ ($L = 1$ then $L = 3$)

(b) $L = 3$ ($L = 1$ then $L = 3$)

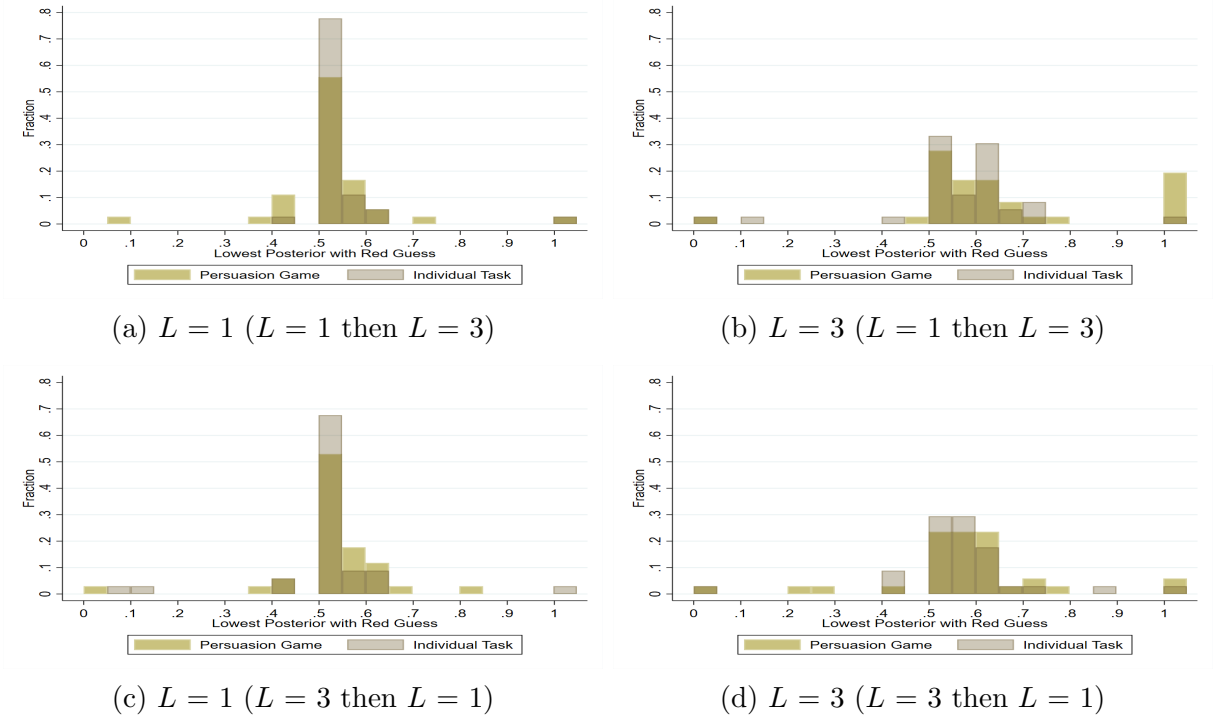
(c) $L = 1$ ($L = 3$ then $L = 1$)

(d) $L = 3$ ($L = 3$ then $L = 1$)

Figure 5 plots the Receivers' aggregate response in the persuasion game with that in the individual task for each value of L. By direct inspection, it is apparent that for both values of L and almost all posterior levels above 0.5, the Receivers guess red more often in the individual task than in the persuasion game. Moreover, this difference is more pronounced

for the case of $L = 3$ than $L = 1$.[15]

Figure 6: Lowest Posterior with Red Guess Comparison (Persuasion Game vs. Individual Task) by L



(a) $L = 1$ ($L = 1$ then $L = 3$)

(b) $L = 3$ ($L = 1$ then $L = 3$)

(c) $L = 1$ ($L = 3$ then $L = 1$)

(d) $L = 3$ ($L = 3$ then $L = 1$)

Figure 6 provides an alternative perspective on the comparison. It contrasts the distribution of the lowest posterior under which a Receiver guesses Red. The figure shows that in both cases of $L = 1$ and $L = 3$, the cutoffs of the Receivers are generally lower in the individual task than in the persuasion game. In other words, the Receivers are generally more demanding in the persuasion game, requiring a stronger posterior to be persuaded. Notably, significantly more Receivers are contented with a posterior in the region $[0.5, 0.55]$ in the individual task than in the persuasion game, especially when $L = 1$.

Table 4 compares the Receivers' behaviors between the individual task and the persuasion game, under the condition that the proportion of red balls in the drawn box is no less than 50%. We focus on this region of posteriors because posteriors below 0.5 almost always fail to persuade. When $L = 1$, conditional on the posterior no less than 0.5, the proportion of red guesses is 81% in the individual task, which is above the 76% in the persuasion game. The difference is significant with p-value equal to 0.04 under two-sample test of proportions. When $L = 3$, the corresponding proportions are 77% and 64% respectively, and the difference is significant with p-value equal to 0.00. The same pattern is preserved when pooling the

[15]The order of L affects neither the frequency of the Red guesses nor the treatment effect significantly.

sample across both values of L, as long as we focus on those boxes with more than 50% of red balls.

Table 4: Comparison of Receiver's Guess in Persuasion Game vs. Individual Task

Guessing Red	Persuasion Game	N	Individual Task	N	Proportion test p-values
Pooled	0.31	2800	0.34	2800	0.01**
	(0.01)		(0.01)		
Pooled & Proportion of Red>0.5	0.70	1181	0.79	1173	0.00***
	(0.01)		(0.01)		
$L = 1$ & Proportion of Red>0.5	0.76	593	0.81	585	0.04**
	(0.02)		(0.02)		
$L = 3$ & Proportion of Red>0.5	0.64	588	0.77	588	0.00***
	(0.02)		(0.02)		

Notes: Significance levels: * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Column 1 of Table 5 reports the probit regression on guessing red, controlling for factors including the proportion of red balls, session, and round effects. The marginal effect coefficients of the probit regression confirm the results of the t-test that subjects are more likely to guess red in the individual task. Columns 2 and 3 report the regression results conditional on the value of L, and the proportion of red balls in the drawn box exceeding 0.5. The results are consistent with the t-tests. Receivers are less inclined to guess red in the persuasion game, and this effect is more significant in the case of $L = 3$. Summarizing the findings above,

Result 4: Receivers are less likely to be persuaded to guess red in the persuasion game than in the individual task. This effect is more significant in the case $L = 3$.

As the Receivers face an identical monetary payoff structure in the individual task and the persuasion game, the difference in behaviors across the two settings can only be attributed to their concerns about the Senders' payoffs. The finding that the Receivers in our experiment react more unfavorably to intermediate posteriors when supplied by the Senders than when drawn by the computer suggests that they may have other-regarding preference. Intuitively, if the ball allocation involves less mixing of red and blue, the realized posteriors are more spread out (with the stronger posterior closer to 1), allowing the Receiver to make a more informed decision and thus enjoy a better expected payoff. This, however, lowers the chance of a red guess, thus hurting the Sender. A more separating ball allocation (equivalently, more revealing signal structure) therefore can be interpreted as a transfer of expected payoff from the Sender to the Receiver. A Receiver may therefore perceive the Sender as hostile if he

Table 5: Determinants of Receiver's Red Guess

	Dependent variable: Guessing Red		
	(1) Pooled	(2) $L = 1$ & Red>0.5	(3) $L = 3$ & Red>0.5
Proportion of Red	1.38***	1.12***	1.49***
	(0.13)	(1.47)	(0.20)
Individual Task	0.05***	0.22*	0.14***
	(0.02)	(0.13)	(0.04)
L Dummy	-0.10***		
	(0.02)		
Session	0.01	0.07*	0.02
	(0.01)	(0.04)	(0.01)
Round	-0.00	0.02	-0.01*
	(0.00)	(0.01)	(0.00)
Number of Balls	-0.00*	-0.01	0.00
	(0.00)	(0.01)	(0.00)
N	5600	1178	1176
Pseudo R2	0.58	0.15	0.20

Notes: This table reports the marginal effect coefficient estimates of the probit regression on the determinants of guessing red. Random-Offer is a dummy that equals 1 when the offer is random, zero otherwise. L dummy equals 1 when L=3, zero otherwise. Significance levels: * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

offers less informative posteriors (than what she is "justly entitled to"), whereas she may perceive the Sender as kind if he offers more informative posteriors. To penalize Senders who are perceived as hostile, the Receiver may opt for a blue guess despite incurring some loss in expected payoff. Receivers who attempt to reciprocate would thus demand a higher cutoff posterior for a red guess than that implied by self-interest calculation. Consequently, compared with their decisions in the individual task, Receivers are less inclined to guess red following less informative posteriors (e.g., those only slightly above 0.5) in the persuasion game.

A crucial element in the theory of reciprocation is the players' perception of their "just" entitlement of payoff. The finding that the Receivers become more demanding from the Senders in the case $L = 3$ may arise because when the stake gets large, the Receiver's payoff becomes more sensitive to the Sender's offer of information.[16] This heightened payoff sensitivity may bring about a larger impact on the Receiver's perception of entitlement, driving up the wedge between the cutoff demanded and the self-interest cutoff. In fact, under

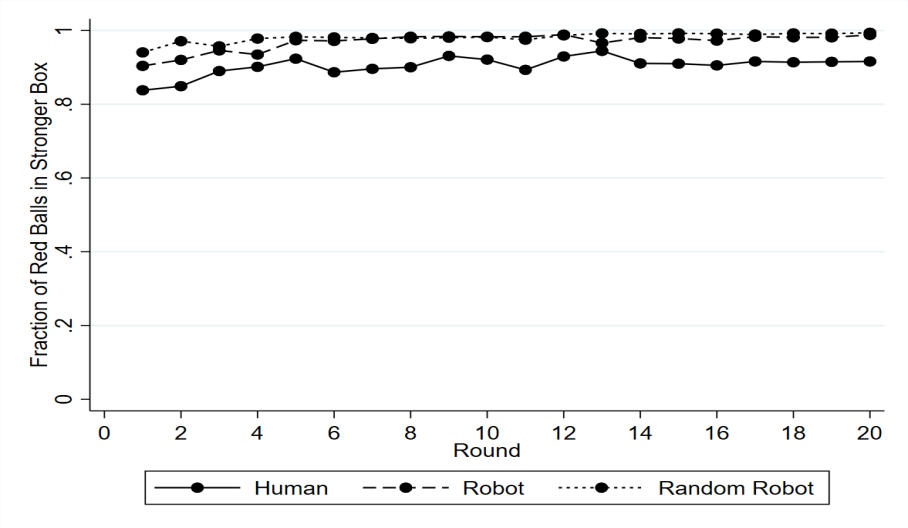
[16]Restricting to the class of signal structure with posteriors 0 and p, a marginal increase in p (corresponding to a more revealing signal structure) has a larger impact on the Receiver's payoff in the case of $L = 3$ than $L = 1$.

the parameter configuration of our experiment, the marginal benefit of a more revealing signal accrued to the Receiver exceeds the marginal cost suffered by the Sender only in the case of $L = 3$. Efficiency-conscious Receivers may therefore become particularly demanding from the Sender when $L = 3$ than when $L = 1$ (e.g., Charness and Rabin, 2002).

3.3 Fully Revealing Weaker Signal

We now turn to the Sender's side. In this section, we look into one of the key theoretical predictions that the Sender sets the weaker signal to be fully revealing (i.e. setting the weaker posterior to be 0), corresponding to putting all the red balls in the stronger box. Starting from this subsection, we will report results from both the Human Receiver treatment and the two Robot Receiver Treatments.

Figure 7: Prop. of Red Balls in the Stronger Urn Over Time



By studying the human Receivers' responses reported in Section 3.1, putting all the red balls in the stronger box is indeed optimal in the Human Receiver treatment, in line with the prescription by the theory of Bayesian persuasion. Figure 7 reveals that, for all treatments, the average proportion of red balls put in the stronger box is high at the beginning and generally increases over time. Recall, given our specification of robots' strategies, it is optimal for the Senders to put all the red balls in the stronger box in both the Robot Receiver treatment and the Random Robot Receiver treatment. Figure 7 reveals that, in this respect, Senders in both the Robot Receiver treatment and the Random Robot Receiver treatment outperform their counterparts in the Human Receiver treatment. This phenomenon suggests that strategic uncertainty from the Receiver side per se can hinder the learning of the optimal

strategy. Nevertheless, even in the Human Receiver treatment, Senders can on average end up putting more than 90% of the Red balls in the stronger box.

Table 6: Prop. of Senders Putting 90% (100%) of the Red Balls in the Stronger Urn

	From Round 1	From Round 6	From Round 11	From Round 16
Treatment 1	24.29% (15.71%)	42.86% (35.71%)	52.86% (47.14%)	62.86% (51.43%)
Treatment 2	64.52% (61.29%)	77.42% (67.74%)	87.10% (67.74%)	90.32% (74.19%)
Treatment 3	71.87% (68.75%)	87.50% (78.13%)	90.62% (87.50%)	93.75% (93.75%)

At the individual level, many Senders choose to stick to putting all the Red balls in one box until the end of the experiment once they learn to do so. Table 6 shows the proportion of Senders who, starting from rounds 1, 6, 11, and 16 (i.e., the beginning of each quarter of the session), persistently put at least 90% (100%) of the red balls in the stronger box until the end of the experiment. Again, we see that the subjects in the two Robot Receiver Treatments generally outperform those in the Human Receiver Treatment, with over 90% of the Senders eventually putting 90% of the Red balls in one box. But even in the Human Receiver Treatment, 15.71% of the Senders put all the red balls in the stronger box in the first round and continue to do so throughout the experiment. More than half of the Senders learn by round 16 to put all the red balls in the stronger urn and stick with this strategy until the end. This suggests that for most Senders, fully revealing the weaker signal is a normatively appealing rule, which is not difficult for them to learn within a session.

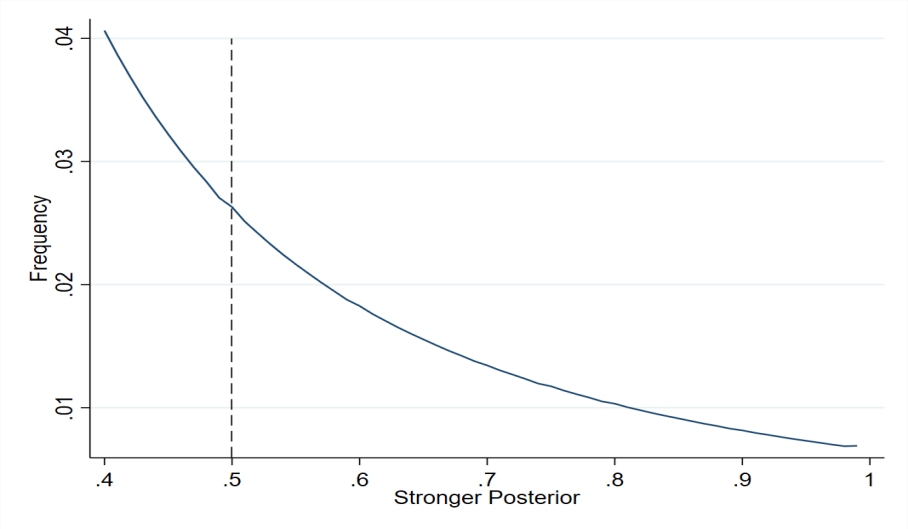
Result 5: The majority of Senders put all the red balls in the stronger box, leaving the weaker box with blue balls only (i.e., the weaker posterior is fully revealing).

3.4 Setting Stronger Posterior Right

We now turn to how the Senders set the stronger posterior. We focus on the stronger posterior because the weaker posterior is necessarily below 50% (recall the prior is below 50%), and is thus generally not enough to persuade the Receiver to guess red. In fact, in our Human Receiver treatment, where the weaker posteriors do not automatically induce a blue guess, only 1.14% of the weaker posteriors induce a red guess.

As the strategy space is large, we first set up a benchmark by asking how a nonstrategic Sender who simply uniformly randomizes over all the available strategies, a 41×61 matrix in our experiment, would set the stronger posterior. Figure 8 shows the (simulated) distribution of the stronger posterior generated by such a uniformly randomizing Sender. It is apparent that posteriors closer to the prior (40%) are more likely to be chosen by a uniformly randomizing Sender. Interestingly, with about a 1/3 probability, a uniformly randomizing

Figure 8: The Distribution of the Stronger Posterior under Uniform Randomization



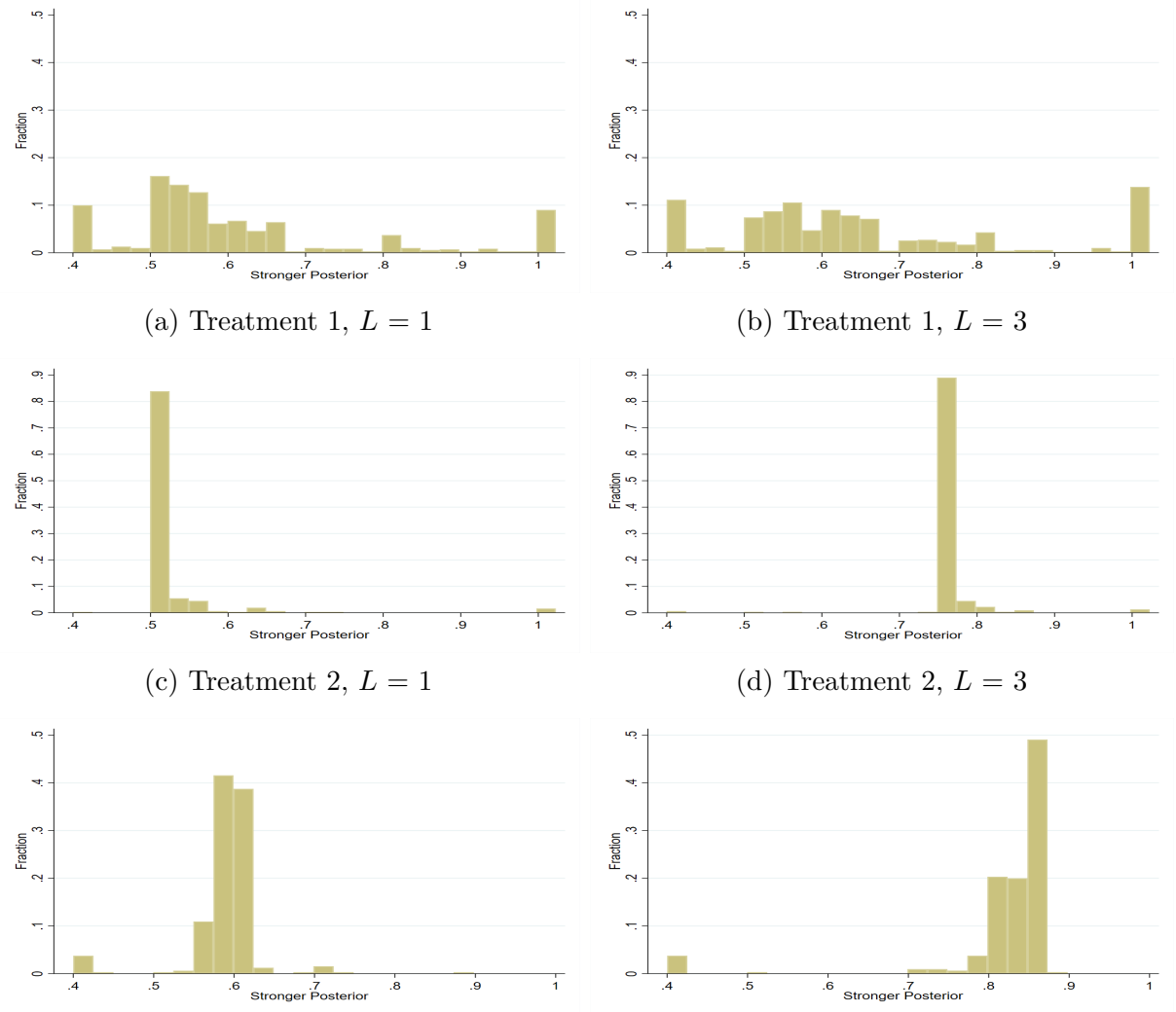
Sender will set the stronger posterior to be lower than 50% —a choice that is very hard to be rationalized by any reasonable Receiver's behavior. Therefore, whether the stronger posterior is higher than 50% can serve as a basic rationality test for the Sender.

Figure 9 reports the empirical distribution of the stronger posterior by treatment and L. The result for the Senders in the Robot Receiver Treatment (Treatment 2), where the Receiver's cutoff is set at $L/(L+1)$, is remarkably stark in that they almost always set the posterior just high enough to persuade. One may argue that this is driven by some anchoring effect as we directly tell the Senders in this treatment the cutoff posterior of the Robot Receiver. However, in the Random Robot Treatment (Treatment 3), where the cutoff is drawn from $U(L/(L+1), L/(L+1)+0.1)$, the Senders generally set the stronger posterior nearly optimally around the top of the distribution, although there is no obvious anchor in this treatment. Thus, the results of our two Robot treatments indicate that the Senders generally understand the importance of setting a high enough posterior to persuade the Receiver to guess Red. They also understand that setting a high posterior has no use beyond inducing the Receivers to guess Red.

It is clear from Figure 9 that in the Human Receiver treatment, where the Receivers' strategies are not defined for the Senders, the stronger posterior is rarely set below 50%, with this event occurring in only 13% (13.57%) of the games when $L = 1$ ($L = 3$).[17] Moreover, the Senders are generally responsive to the changes in the value of L, as they generally set

[17]Frechette et al. (2022) find that a significant proportion of the Senders in their experiment set the stronger posterior to be lower than 50% even with ample experience. It is another indication that the subjects' failure of Bayesian updating plays an important role in their environment.

Figure 9: Empirical Distribution of Stronger Posterior by Treatment and L



(a) Treatment 1, $L = 1$

(b) Treatment 1, $L = 3$

(c) Treatment 2, $L = 1$

(d) Treatment 2, $L = 3$

(e) Treatment 3, $L = 1$

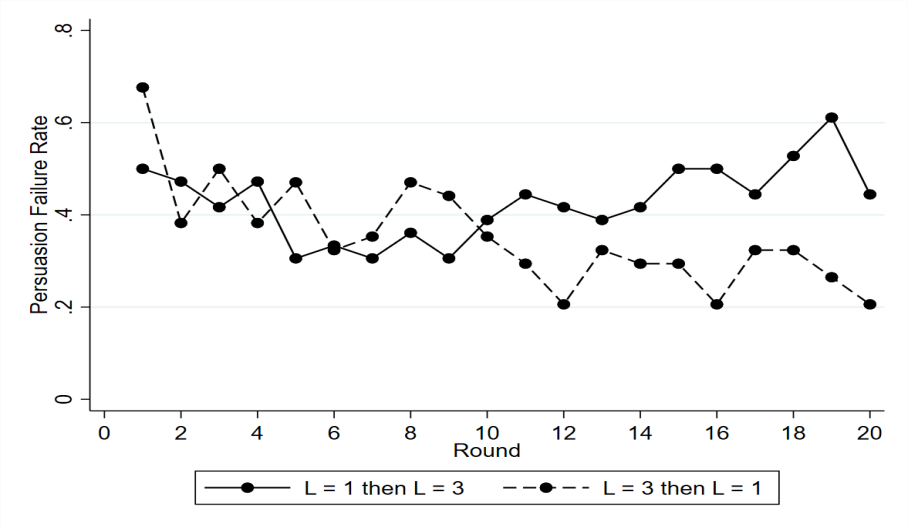
(f) Treatment 3, $L = 3$

the stronger posterior higher under higher L ($p < 0.01$, Wilcoxon rank sum test).

However, the Senders are generally not responsive enough to changes in the value of L. To see this, recall that, from Figure 2, to persuade most Human Receivers to guess Red, the stronger posterior needs to be higher than 50% when $L = 1$ and 60% when $L = 3$. While only 13% of Senders set posteriors below 50% when $L = 1$, about 45% of them set posteriors below 60% when $L = 3$. Moreover, although some Senders set the stronger posterior at 100%, they rarely set the stronger posterior within [80%, 100%) when $L = 3$, although in terms of payoff, setting the stronger posterior in this region is close to the empirical best response.

Importantly, there is only weak evidence that Senders increase the stronger posterior as they gain experience. To see this, we compare the stronger posterior set in the first five

Figure 10: Persuasion Failure Rate Over Time in Human Receiver Treatment



rounds and in the last five rounds under each L. We find that, on average, the Senders increase the average posterior by 0.02 when $L = 1$ ($p > 0.1$) and 0.03 when $L = 3$ ($p < 0.05$) (two-sided t-test, clustered at the sender level). The magnitude of learning on the Sender side in terms of setting the stronger posterior is thus negligible.

This systematic underprovision of information leads to frequent and persistent *persuasion failure* (i.e., Blue guess under even the stronger signal). This phenomenon is striking. Recall Figure 4 reveals that under-informing Receivers is a more costly mistake than over-informing Receivers. In fact, When $L = 3$, Senders can achieve a higher (ex-ante) persuasion rate by completely separating the balls into the two boxes (i.e., full revelation) than by setting the stronger posterior below 60%.

Figure 10 shows how the persuasion failure rate changes over time. Clearly, the persuasion failure rate generally stays high throughout the whole session. Moreover, the higher the L, the more frequent the persuasion failure: while 33.14% of the stronger posteriors induce a Blue guess when $L = 1$, 45.29% of the stronger posteriors induce a Blue guess when $L = 3$. This is in sharp contrast to the equilibrium prediction that the stronger signal always induces a Red guess, as well as the two Robot Receiver treatments, where persuasion failure is much rarer (1.13% in Robot Receiver treatment and 19.37% in Random Robot Receiver treatment).

Result 6: Senders generally respond to the change of L. The vast majority of them set the stronger posteriors either within [50%, 80%] or at 100%. They rarely set the stronger posterior in (80%, 100%] even when $L = 3$. The stronger posterior is often too low for persuasion, resulting in persistently frequent persuasion failure.

In light of the fact that most Senders in the Random Robot treatment seem to understand

that under-informing Receivers is a more costly mistake than over-informing Receivers, the persistently high rate of persuasion failure, especially in the case of $L = 3$, is puzzling.[18] First, we find that the high persuasion failure rate is unlikely to be driven by those confused Senders who are unable to properly understand the tasks at hand. The high persuasion failure is a rather systemic pattern, as 40% of the Senders have more than half of their stronger posteriors inducing a Blue guess. Moreover, the same behavioral pattern is also found in one of our previous drafts (Kwon, 2020) with a different subject pool. Second, the social preference of the Senders cannot explain the under-informing behavior of the Senders. This is because setting the strong posterior too low for successful persuasion hurts the welfare of both the Sender and the Receiver. Third, random decision error (e.g. Quantal Response) cannot explain the behavioral pattern of the Senders because on one hand, they do not set the posterior below 50%, whereas on the other hand, neither do they set the posterior within (80%, 100%) even when doing so is close to the empirical best response[19].

We turn to a fourth explanation —Sender's optimism about the Receiver's cutoff. From the Robot Receiver treatment, we see that the Senders are generally aware that the optimal strategy is to match the Receiver's cutoff. Thus, if a Sender holds a mistaken belief that the Receiver has a lower cutoff than she actually has, he may set his stronger posterior too low, resulting in persuasion failure.

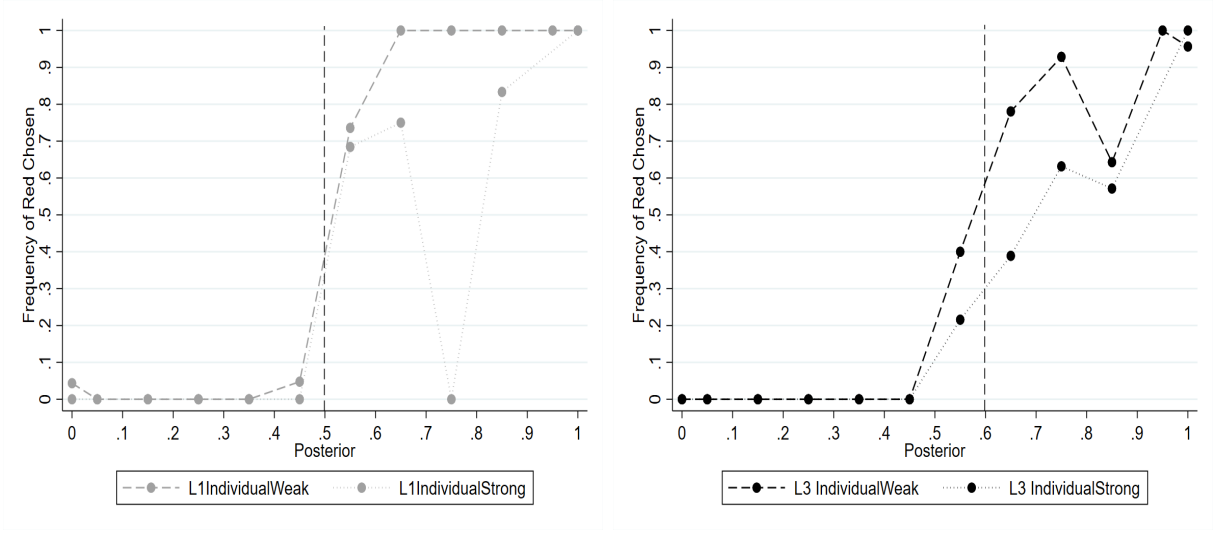
To formally test this explanation, we look at the Senders' belief about the Receivers' guesses. Recall that in our Human Receiver treatment, after the persuasion game, the Senders are given an individual task in which they are asked to guess whether the Receiver chose Red in some past rounds of the persuasion game. A monetary prize is awarded if and only if their guess matches the Receiver's actual guess in that round. Thus, the Sender's guess is an indicator of whether they believe that a given posterior can induce *most Receivers* to guess Red.

Figure 11 compares, for each L, the guesses of those Senders whose average stronger posterior is above the median level ("the strong Senders") and those below the median level ("the weak Senders"). As most Receivers actually guess Red when the posterior is higher than 50% when $L = 1$ and 60% when $L = 3$, the optimal guess of the Senders is guessing Red when the posterior is higher than 50% under $L = 1$ and 60% under $L = 3$. While most Senders guess optimally under $L = 1$, the strong Senders are generally more pessimistic about the Receivers' response. A logit regression for the case $L = 3$ confirms that holding the posterior probability constant, the strong Senders are less likely to guess that the Receiver

[18]Note that any explanation related to the Bayesian updating is automatically ruled out because the subjects do not bear the burden of calculating the posteriors.

[19]See Appendix A for full description of our Quantal Response Equilibrium (QRE) estimation.

Figure 11: Individual Decision Comparison between Strong and Weak Senders



(a) L = 1 (b) L = 3

chose Red ($p < 0.1$, clustered at the Sender level). This suggests that optimism can partially explain why some Senders set the stronger posterior too weak.

4 Discussion

In summary, we find that the Senders can understand the normative appeal of fully revealing the weaker signal, but they systematically set the stronger posterior lower than what the Receivers require, leading to frequent persuasion failure even under the stronger signal. The frequent persuasion failure is puzzling because the Sender can always avoid this costly mistake by providing full information. Yet, it is a very robust phenomenon across subjects. But when the Senders know the Receiver's strategy exactly, the Sender can generally choose the optimal strategy.

Although Bayesian persuasion is relatively abstract, our results suggest that its basic strategic element, when cleanly isolated, can be easily understood by our human subjects who presumably have little relevant knowledge. Our experimental design, which interprets Bayesian persuasion in terms of information partition, makes such clean isolation possible while keeping the problem interpretable. As a result, we are able to obtain sharp results and identify some important behavioral patterns that would otherwise be difficult to detect.

Our experimental design allows for a variety of extensions based on our knowledge of the information partition. For example, we know that we can combine two pieces of information by taking the join of their respective information partition. Consequently, one way to

implement competitive Bayesian Persuasion between multiple Senders is to let each Sender choose a partition and then give the Receiver the join of the partitions. However, it must be noted that our design is only applicable when the Receiver is supposed to know the posterior probability exactly. For example, we cannot replicate some of the treatments of Frechette et al. (2022) because in those treatments, the interpretation of the signals is subject to the Sender's manipulation, and thus a naive Receiver may not know the posterior probability exactly.

Although our design is minimal and may exclude some interesting behavioral aspects, we believe that as long as the strategic element under study is orthogonal to those aspects, a minimal design like ours provides a useful starting point to understand the extent to which the strategic element is alive. Here, we find that the strategic element of Bayesian Persuasion is not only alive but very strong, and we expect it will also be strong in many other extensions.

References

Aristidou, A., G. Coricelli, A. Vostroknutov, et al. (2019). *Incentives or persuasion? an experimental investigation*. Maastricht University, Graduate School of Business and Economics.

Au, P. H. and K. K. Li (2018). Bayesian persuasion and reciprocity: theory and experiment. *Available at SSRN 3191203*.

Blume, A., E. K. Lai, and W. Lim (2020). Strategic information transmission: A survey of experiments and theoretical foundations. *Handbook of Experimental Game Theory*.

Boleslavsky, R. and C. Cotton (2015). Grading standards and education quality. *American Economic Journal: Microeconomics 7*(2), 248–79.

Brooks, B., A. P. Frankel, and E. Kamenica (2019). Information hierarchies. *Available at SSRN 3448870*.

Charness, G. and M. Rabin (2002). Understanding social preferences with simple tests. *The quarterly journal of economics 117*(3), 817–869.

Dickhaut, J. W., K. A. McCabe, and A. Mukherji (1995). An experimental study of strategic information transmission. *Economic Theory 6*(3), 389–403.

Eckel, C. C. and P. J. Grossman (2008). Forecasting risk attitudes: An experimental study using actual and forecast gamble choices. *Journal of Economic Behavior & Organization 68*(1), 1–17.

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*(3), 481–518.

Goeree, J. K., C. A. Holt, and T. R. Palfrey (2016). Quantal response equilibrium. In *Quantal Response Equilibrium*. Princeton University Press.

Green, J. R. and N. Stokey (1978). *Two representations of information structures and their comparisons*. Institute for Mathematical Studies in the Social Sciences.

Hagenbach, J. and E. Perez-Richet (2018). Communication with evidence in the lab. *Games and Economic Behavior 112*, 139–165.

Hsee, C. K. and E. U. Weber (1997). A fundamental prediction error: Self–others discrepancies in risk preference. *Journal of experimental psychology: general 126*(1), 45.

Hsee, C. K. and E. U. Weber (1999). Cross-national differences in risk preference and lay predictions. *Journal of Behavioral Decision Making 12*(2), 165–179.

Jin, G. Z., M. Luca, and D. Martin (2017). Is no news (perceived as) bad news? an experimental investigation of information disclosure. Technical report, Harvard Business School.

Kamenica, E. and M. Gentzkow (2011). Bayesian persuasion. *American Economic Review 101*(6), 2590–2615.

Kwon, O. (2020). Bayesian persuasion in the lab. *Working Paper*.

Meng, D. and S. Wang (2022). Information avoidance from exogenous vs endogenous sources: Theory and experiment. *Available at SSRN*.

Siegrist, M., G. Cvetkovich, and H. Gutscher (2002). Risk preference predictions and gender stereotypes. *Organizational Behavior and Human Decision Processes 87*(1), 91–102.

Wu, W. and B. Ye (2022). Competition in persuasion: An experiment. *Games and Economic Behavior*.

Appendix A: Quantal Response Equilibrium (QRE)

In this section of Appendix, we present the results of Quantal Response Equilibrum (QRE) estimation of our persuasion game (see, e.g., Goeree et al., 2016). We will show that although the best-fitting QRE can capture some aspects of the observed behavior, it cannot explain some important behavioral patterns that our experiment reveals.

We first lay down some definitions needed for the estimation. We define a Sender's strategy as $\pi = (p_{str}, q_{str}) \in \Pi$, where p_{str} and q_{str} denote the frequency of the stronger box and the fraction of red balls in the stronger box, respectively. We define a Receiver's strategy as $\sigma(\pi) = (\sigma_{str}, \sigma_{weak})(\pi) \in \Sigma$, where σ_{str} and σ_{weak} denotes whether the Receiver guesses Red under the stronger and the weaker posterior respectively.

In QRE, each player may make mistake in his or her decisions but has a correct belief about the other player's mistake. The size of the Sender's and the Receiver's mistake is quantified by λ_s and λ_r respectively, which are commonly known between the Sender and the Receiver. However, because the Sender moves first, from the Receiver's perspective, the Sender's decision error is irrelevant to her decision. Thus, the sequential structure of the persuasion game allows us to estimate λ_s and λ_r separately.

QRE assumes that the probability that the Receiver plays each strategy is

$$Pr(\sigma|\pi, \lambda_r, L) = \frac{\exp\left(\lambda_r U(\sigma|\pi, L)\right)}{\sum_{\sigma'} \exp\left(\lambda_r U(\sigma'|\pi, L)\right)}$$

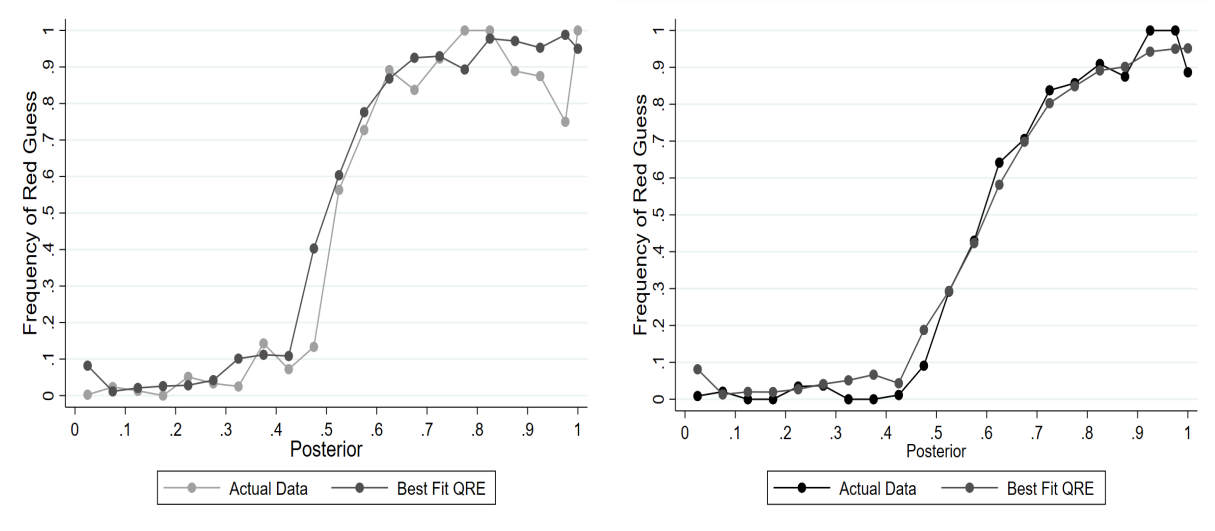
where $U(\sigma|\pi, L)$ denotes Receiver's ex-ante utility when she follows strategy σ when the Sender follows strategy π under parameter L. Clearly, the probability that the Receiver plays the optimal strategy increases in λ_r, and thus this parameter captures how unlikely the Receiver makes mistake.

Given the expression of the probability that the Receiver uses each strategy, we use the maximum likelihood estimation to find $\hat{\lambda_r}$ that best fits the empirical distribution. Importantly, the Receiver's payoff is strongly affected by her risk attitude when $L = 3$, thus, instead of imposing risk neutrality on the Receiver, we also estimate $\hat{L}$ that maximizes the likelihood function when $L = 3$. It is easy to show that the likelihood function is concave in both $\hat{\lambda_r}$ and $\hat{L}$ so that there is a unique maximizer of the likelihood function.

Figure 12 compares how the Receivers respond to each level of posterior under the best fitting QRE and in the data under each L. We find that QRE fits the data extremely well in most regions.[20] However, when $L = 1$, when the posterior is between 45% and 50%, the best fitting QRE predicts that the red guess is chosen approximately 40% of the times but in

[20] When $L = 1$, we have $\hat{\lambda_r} = 13.52$. When $L = 3$, we have $\hat{\lambda_r} = 7.51$ and $\hat{L} = 1.47$

Figure 12: Receiver Response Comparison between Actual Data and Best Fitting QRE



(a) L = 1 (b) L = 3

the actual data it is chosen only approximately 10% of the times[21]. A key prediction of the QRE model in our persuasion game when $L = 1$ is that the Receiver's response function is approximately symmetric around 50% because the payoff from guessing Red under a slightly lower than 50% posterior is symmetric to that from guessing Blue under a slightly higher than 50%. Thus, the QRE model cannot explain why the Receiver makes one type of mistake significantly more frequently than the other.

When $L = 3$, after accounting for risk aversion, we find that the QRE fits the data well. Importantly, the best fitting $\hat{L}$ is 1.47, which is significantly less than the treatment variable L. Note that the threshold likelihood ratio of $\hat{L} = 1.47$ corresponds to the threshold posterior of 59.5%, which is close to what we find as the threshold above which most Receiver guesses Red. For reference, when we impose $\hat{L}$ to be 3, we find that the best fitting QRE has a much lower $\hat{\lambda_r}$[22], meaning that if we impose an incorrect assumption on the Receiver's risk attitude, the model requires the Receiver to be significantly more irrational in order to best fit the data.

Now that we have estimated the best fitting QRE on the Receiver's side, we can now turn to the estimation of λ_s, the Sender's parameter. QRE assumes that the probability that

[21]In our data, the posterior slightly higher than 50% is much more frequent than the posterior slightly lower than 50%. This is why the estimation puts more weight on the former region rather than the latter.

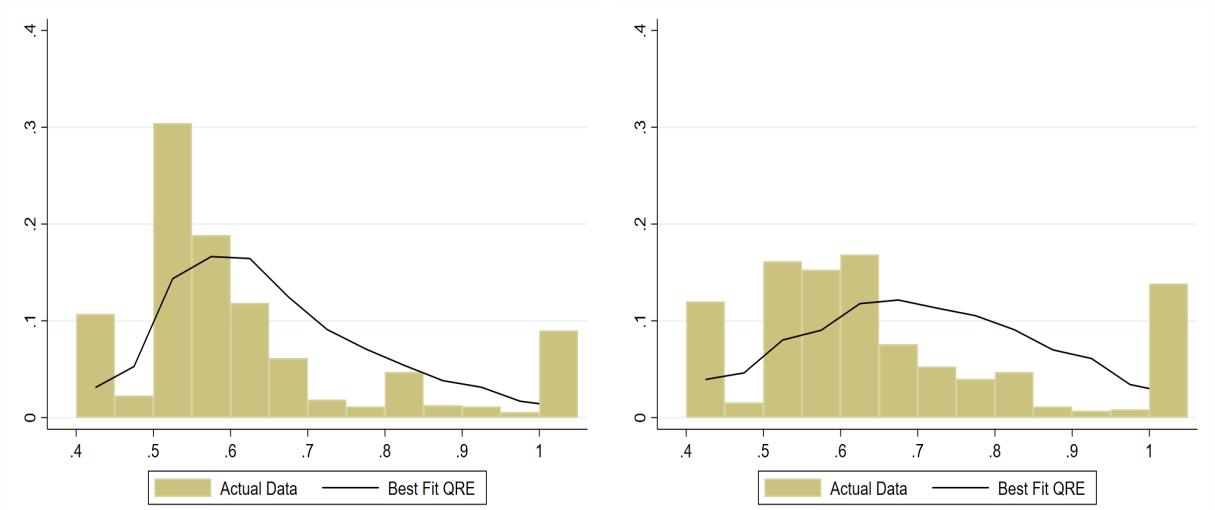
[22]When we impose $\hat{L} = L = 3$, we have $\hat{\lambda_r} = 2.75$.

the Sender plays each strategy is

$$Pr(\pi|\hat{\lambda_r}, \lambda_s, L) = \frac{\exp\left(\lambda_s V(\pi|\hat{\lambda_r}, \lambda_s, L)\right)}{\sum_{\pi'} \exp\left(\lambda_s V(\pi'|\hat{\lambda_r}, \lambda_s, L)\right)}$$

where $V(\pi|\hat{\lambda_r}, \lambda_s, L)$ denotes the Sender's payoff when he follows strategy π when the Receiver is of type $\hat{\lambda_r}$ under parameter L. Similar to the Receiver, the probability that the Sender plays the optimal strategy increases in λ_s.

Figure 13: Sender's Stronger Posterior Comparison between Actual Data and Best Fitting QRE



(a) L = 1 (b) L = 3

We again use the maximum likelihood estimation to find λ_s that best fits the empirical distribution of the Sender's strategy. Figure 13 compares how the Sender sets the stronger posterior under the best fitting QRE and in the data under each L. We find that, regardless of L, QRE generally predicts a higher frequency of high posteriors than the actual data. In particular, compared to the actual data, QRE puts significantly more weight on the posteriors within the $(80\%, 100\%)$ region. This is yet another reflection of the fact that over-informing is generally a less costly mistake than under-informing. Thus, QRE model cannot explain why the Sender sets the posterior systematically lower than what is optimal.

Appendix B: Instructions

Instructions for Treatment 1 (Human Receiver Treatment)

Welcome and thank you for participating in this study of economic decision-making. We expect the study to take about 90 minutes.

The instruction is simple. If you follow carefully and make good decisions, you may earn a substantial amount of money.

This experiment consists of two stages. We will explain and provide instruction of each stage when we reach there. Your choices in any one stage will not affect what happens to you in the other. Therefore, we recommend that you focus on the current stage without thinking about the later stage.

Stage 1

In this stage, you are assigned either Role 1 or Role 2, which remains fixed throughout the study. A Role 1 will be randomly matched with a Role 2 in each round. Therefore, if you are Role 1, it is unlikely that you are matched with the same Role 2 of the previous rounds, and vice versa.

There will be a total of 20 rounds in this stage.

Role 2 needs to guess the color of a ball drawn from a box that contains Red and Blue balls. If the guess is correct, Role 2 gets a positive payoff; otherwise, Role 2 gets no payoff.

Role 1 needs to provide hints to convince Role 2 to guess Red. If Role 2's guess is Red, no matter whether it is correct or not, Role 1 gets a positive payoff; otherwise, Role 1 gets no payoff.

Specifically, each round proceeds as follows.

The Drawn Ball

A can contains 100 balls labeled 1, 2, . . . , 100. Out of these 100 balls, **40 balls are Red and 60 balls are Blue**.

The computer randomly draws a ball, records its label (e.g., 72), and puts it back into the can. Note that every ball is equally likely to be drawn.

Throughout the decision-making stages, we will not tell anybody anything about the drawn ball. In the rest of the instruction, we will call this ball "*the drawn ball*."

Role 1's Decision

Role 1 moves first. Role 1 decides how to divide the 100 balls between two empty boxes, Box A and Box B. Role 1 is free to divide the 100 balls between these two empty boxes in any way.

To do this, Role 1 can simply indicate the number of Red balls and Blue balls he/she wants to put in Box A, and the computer will do the rest.

Example: Role 1 can place 15 Red balls and 20 Blue balls in Box A. The computer will then automatically put the rest of the balls (25 Red balls and 40 Blue balls) into Box B. The final composition of the boxes will be as follows:

Box A: 15 Red Balls and 20 Blue Balls (35 balls in total)
Box B: 25 Red Balls and 40 Blue Balls (65 balls in total)

The following facts are worth noting.

- Role 1 does not know the color of the drawn ball.
- Role 1 cannot affect the overall chance that the drawn ball is a Red one (40% = 40/100), as the drawn ball was determined by the computer before Role 1 is asked to act.
- One of the boxes, Box A or Box B, must contain the drawn ball no matter how Role 1 divides the 100 balls.
- As each ball is equally likely to be the drawn ball, the more balls a box contains, the more likely that it is holding the drawn ball. In fact, the chance that a box holds the drawn ball is proportional to the total number of balls placed inside. For example, if there is a total of 35 balls in Box A, then the chance that it holds the drawn ball is 35%. In this case, the chance that Box B holds the drawn ball is 65%.
- As the computer has recorded the label of the drawn ball, it can tell whether the drawn ball is contained in Box A or Box B (but no one else can tell exactly).

Role 2's Decision

After Role 1 completes the ball division, it is Role 2's turn to make decisions.

The computer will give Role 2 the box that contains the drawn ball (which may either be Box A or Box B), and reveal its composition to Role 2. The other box becomes irrelevant for the payoffs of both Roles, and is effectively thrown away.

The task of Role 2 is to form a guess about the color of the drawn ball, based on the two pieces of facts below.

- The drawn ball is contained in the box received.
- The numbers of Red and Blue balls in the received box are as shown by the computer.

After Role 2 has made the guess, the computer then compares it with the actual color of the drawn ball, and computes the payoff of each role as follows.

- Role 1 gets a positive payoff if Role 2's guess is Red.
- Role 2 gets a positive payoff if Role 2's guess is correct.

In the experiment, before Role 2 actually receives his/her box, we ask Role 2 to form a plan of guesses. That is, *for both Box A and Box B, Role 2 needs to make a guess on the color of the drawn ball* ***supposing that*** *the drawn ball is contained in that box.* Specifically, Role 2 is asked both of the following questions:

- Suppose you receive Box A, so that Box A contains the drawn ball. What is your guess on the color of the drawn ball? —— Red/Blue
- Suppose you receive Box B, so that Box B contains the drawn ball. What is your guess on the color of the drawn ball? —— Red/Blue

(If Role 1 does not put any ball in one of the boxes, the question for that box will be skipped.)

As only one box can actually contain the drawn ball and only that box will actually be given to Role 2, we will take Role 2's guess for the received box as her actual guess.

As Role 2, when making the plan of guesses, does not know whether Box A or Box B will eventually be received, the answer to each question above can potentially become the actual guess, and hence relevant to the payoffs of the matched pair. We therefore recommend that Role 2 makes the guess for each box ***as if*** it is the box to be received.

The details of payoff determination are explained below.

Payoffs

Role 1 gets **70 points** whenever Role 2's actual guess is Red, and 0 points otherwise. Role 1's payoff points *does not depend on* whether Role 2's guess is correct or not.

Role 2 gets **positive payoff points** if the actual guess is **the same as the color of the drawn ball**, and 0 points otherwise. Specifically, the payoff points that Role 2 receives are determined by the following table.

	If Drawn Ball is Red	If Drawn Ball is Blue
Guessing Red	30	0
Guessing Blue	0	X

In words,

- If the guess is Red, and the drawn ball is Red, Role 2 gets 30 points.
- If the guess is Red, and the drawn ball is Blue, Role 2 gets 0 points.
- If the guess is Blue, and the drawn ball is Red, Role 2 gets 0 points.
- If the guess is Blue, and the drawn ball is Blue, Role 2 gets X points.

The number X in the table is 90 points for the first 10 rounds, and 30 points for the last 10 rounds.

An Example

Suppose Role 1 divides the balls as follows.

Box A: 15 Red Balls and 20 Blue Balls (35 balls in total)
Box B: 25 Red Balls and 40 Blue Balls (65 balls in total)

Box A has 35 balls and Box B has 65 balls. Therefore there is a 35% chance that the drawn ball is in Box A, and a 65% chance that the drawn ball is in Box B.

Let's say Role 2 receives Box A (which happens with a 35% chance). The fraction of Red balls in Box A is 42.86% (=15/35), and the fraction of Blue balls is 57.14% (=20/35).

Having observed the composition of the box, Role 2 decides between guessing Red or Blue.

Suppose Role 2's guess is Red. Then with a 42.86% chance, the guess is correct and Role 2 receives a positive payoff. With a 57.14% chance, the guess is incorrect and Role 2 receives no payoff. With Role 2's guess being Red, Role 1 gets a positive payoff of 70 points with certainty.

Suppose Role 2's guess is Blue. Then with a 57.14% chance, the guess is correct and Role 2 receives a positive payoff. With a 42.86% chance, the guess is incorrect and Role 2 receives no payoff. With Role 2's guess being Blue, Role 1 gets no payoff with certainty.

Summary

1. A ball is randomly drawn from a total of 40 Red and 60 Blue balls, but the identity of the drawn ball is known only to the computer.

2. Role 1 is rewarded if and only if Role 2 makes an actual guess of Red.

3. Role 2 is rewarded if and only if the guess matches the color of the drawn ball.

4. In the first step, Role 1 divides the 100 balls between Box A and Box B, without knowing the color of the drawn ball.

    - From the point of view of Role 1, the more balls a box has, the more likely that the drawn ball is contained inside, and hence the more likely that it will be given to Role 2. In fact, the chance that a box is received by Role 2 is proportional to the number of balls contained inside.

5. With the knowledge of the color composition of the received box, Role 2 makes a guess about the color of the drawn ball.

    - From the point of view of Role 2, every ball in the received box is equally likely to be the drawn ball. Therefore, the chance that the drawn ball is Red (Blue) is equal to the fraction of Red (Blue) balls in the received box.

Monetary Payment

The monetary payment you receive for this stage is determined as follows. The computer will randomly pick one round from the first 10 rounds, and another round from the latter 10 rounds. The sum of the payoff points that you received in these two randomly-selected rounds will be converted to money at an exchange rate of 1 point = 1 HKD, and paid to you. As every round may count, we recommend that you act in each round *as if* it is a round that counts.

Stage 2

In this stage, your role remains the same as the previous stage. That is, if you were Role 1 before, you will remain Role 1 in this stage. If you were Role 2 before, you will remain Role 2 in this stage. However, in this stage, there will be no interaction between Role 1 and Role 2. Each role will perform a separate task independently.
There will be a total of 20 rounds in this stage.

Role 1's Task

In each round, Role 1's task is to guess the choice of a Role 2 in a randomly-selected round in the previous stage. Role 1 can receive a positive payoff if and only if **the guess about Role 2's choice is correct**. More specifically, if you are Role 1, each round proceeds as follows.

1. The computer will randomly pick one round from the previous stage that was played by someone else. Therefore, you were not involved in that past round, and you could not possibly affect what happened in that past round.

2. The computer will show you the Box A and Box B prepared by Role 1 in that past round of the previous stage. Your task is to guess what Role 2 in that round has chosen.

3. Only your guess for the box with the drawn ball counts. As you do not know which box contains the drawn ball, we recommend that you make the guess as if it is the box that counts. If your guess about Role 2's choice is correct, you get a payoff of 40 points; otherwise, you get no payoff.

4. Unlike the previous stage, your decision in this stage has no effect on other participants in any way.

Role 2's Task

Role 2's task in this stage is similar to that in the previous stage with one notable difference – **the decisions made will not affect any other participants in any way**. More specifically, if you are a Role 2, each round proceeds as follows.

1. The computer will randomly pick one round from the previous stage that was played by someone else. Therefore, you were not involved in that past round, and you could not possibly affect what happened in that past round.

2. The computer will set up the Box A and Box B as prepared by Role 1 in that past round of the previous stage. It will also randomly pick a drawn ball from the 100 balls (40 Red balls and 60 Blue balls), which may either be contained in Box A or Box B.

3. The computer will give you whichever box that contains the drawn ball, and ask you to guess its color. Your payoff is determined by the same table as the previous stage.

	If Drawn Ball is Red	If Drawn Ball is Blue
Guessing Red	30	0
Guessing Blue	0	X

   The number X in the table is 90 points for the first 10 rounds, and 30 points for the last 10 rounds.

4. As in the previous stage, before you know which box you actually receive, we require you to guess the color of the drawn ball for both possible cases. That is, for both Box A and Box B, Role 2 needs to make a guess on the color of the drawn ball supposing that the drawn ball is in that box. As only one box can actually contain the drawn ball, we will take your guess for that box as your actual guess.

5. Unlike the previous stage, your decision in this stage has no effect on other participants in any way.

Monetary Payment

The monetary payment you receive for this stage is determined as follows. The computer will randomly pick one round from the first 10 rounds, and another round from the latter 10 rounds. The sum of the payoff points that you received in these two randomly-selected rounds will be converted to money at an exchange rate of 1 point = 1 HKD, and paid to you. As every round may count, we recommend that you act in each round *as if* it is a round that counts.

Instructions for Treatment 3 (Random Robot Receiver Treatment)

Welcome and thank you for participating in this study of economic decision-making. We expect the study to take about 70 minutes.

The instruction is simple. If you follow carefully and make good decisions, you may earn a substantial amount of money.

In this experiment, a robot makes a guess on the color of a ball drawn from a can filled with Red and Blue balls.

Your task is to provide hints to convince the robot to guess that the color of the drawn ball is Red. If the robot's guess is Red, no matter whether it is correct or not, you get a **positive payoff**; otherwise, you get no payoff.

There are 20 rounds in this experiment. Specifically, each round goes as follows.

The Drawn Ball

A can contains 100 balls labeled 1, 2, . . . , 100. Out of these 100 balls, **40 balls are Red and 60 balls are Blue**.

The computer randomly draws a ball, records its label (e.g., 72), and puts it back into the can. Note that every ball is equally likely to be drawn.

Throughout the decision-making stages, we will not tell anybody anything about the drawn ball. In the rest of the instruction, we will call this ball "*the drawn ball*."

Your Decision

You move first. You decide how to divide the 100 balls between two empty boxes, Box A and Box B. You are free to divide the 100 balls between these two empty boxes in any way.

To do this, you can simply indicate the number of Red balls and Blue balls you want to put in Box A, and the computer will do the rest.

Example: You can place 15 Red balls and 20 Blue balls in Box A. The computer will then automatically put the rest of the balls (25 Red balls and 40 Blue balls) into Box B. The final composition of the boxes will be as follows:

Box A: 15 Red Balls and 20 Blue Balls (35 balls in total)
Box B: 25 Red Balls and 40 Blue Balls (65 balls in total)

The following facts are worth noting.

- You do not know the color of the drawn ball.

- You cannot affect the overall chance that the drawn ball is a Red one (40% = 40/100), as the drawn ball was determined by the computer before you are asked to act.
- One of the boxes, Box A or Box B, must contain the drawn ball no matter how you divide the 100 balls.
- As each ball is equally likely to be the drawn ball, the more balls a box contains, the more likely that it is holding the drawn ball. In fact, the chance that a box holds the drawn ball is proportional to the total number of balls placed inside. For example, if there is a total of 35 balls in Box A, then the chance that it holds the drawn ball is 35%. In this case, the chance that Box B holds the drawn ball is 65%.
- As the computer has recorded the label of the drawn ball, the computer can tell whether the drawn ball is contained in Box A or Box B (but no one else can tell exactly).

Robot's Decision

After you complete the ball division, it is the robot's turn to make decisions.

The computer will give the robot the box that contains the drawn ball (which may either be Box A or Box B), and reveal its composition to the robot. The other box becomes irrelevant, and is effectively thrown away.

The robot then forms a guess about the color of the drawn ball, based on the two pieces of facts below.

- The drawn ball is contained in the box received.
- The numbers of Red and Blue balls in the received box are as shown by the computer.

The robot's decision follows a fixed rule as follows.

If Fraction of Red Balls Larger than X,	Robot guesses Red
If Fraction of Red Balls Smaller than X,	Robot guesses Blue
If Fraction of Red Balls Equal to X,	Robot randomizes

In other words,

- When the fraction of Red balls in the box received is larger than X, the robot guesses Red.
- When the fraction of Red balls in the box received is smaller than X, the robot guesses Blue.

- When the fraction of Red balls in the box received is equal to X, the robot guesses randomly.

In the first 10 rounds, the number X is **between 50% and 60%,** with *each value in the interval equally likely.* The number X is drawn afresh in each round.

In the latter 10 rounds, the number X is **between 75% and 85%,** with *each value in the interval equally likely.* The number X is drawn afresh in each round.

Payoffs

You get a payoff of **80 points** whenever the robot's guess is Red, and **0 points** otherwise. Your payoff points *do not depend* on whether the robot's guess is correct or not.

Examples

Example 1: Suppose you divide the balls as follows.

Box A: 15 Red Balls and 20 Blue Balls (35 balls in total)
Box B: 25 Red Balls and 40 Blue Balls (65 balls in total)

Box A has 35 balls and Box B has 65 balls. Therefore there is a 35% chance that the drawn ball is in Box A, and a 65% chance that the drawn ball is in Box B.

Let's say the robot receives Box A (which happens with a 35% chance). The fraction of Red balls in Box A is 42.86% (=15/35), and the fraction of Blue balls is 57.14% (=20/35).

Having observed the composition of the box, the robot decides between guessing Red or Blue. As the fraction of Red balls in Box A is 42.86%, the robot guesses Blue in both cases of X between 50%-60% and 75%-85%.

Example 2: Suppose you divide the balls as follows.

Box A: 22 Red Balls and 18 Blue Balls (40 balls in total)
Box B: 18 Red Balls and 42 Blue Balls (60 balls in total)

Box A has 40 balls and Box B has 60 balls. Therefore there is a 40% chance that the drawn ball is in Box A, and a 60% chance that the drawn ball is in Box B.

Let's say the robot receives Box A (which happens with a 40% chance). The fraction of Red balls in Box A is 55% (=22/40), and the fraction of Blue balls is 45% (=18/40).

Having observed the composition of the box, the robot decides between guessing Red or Blue. As the fraction of Red balls in Box A is 55%, the robot guesses Red if its X value happens to be between 50% and 55%. This happens with a chance about one-half in the first 10 rounds (which has X between 50% - 60%), and zero chance in the last 10 rounds (which has X between 75% - 85%).

Summary

- A ball is randomly drawn from a total of 40 Red and 60 Blue balls, but the identity of the drawn ball is known only to the computer.
- You are rewarded if and only if the robot makes a guess of Red.
- In the first step, you divide the 100 balls between Box A and Box B, without knowing the color of the drawn ball.
- From your point of view, the more balls a box has, the more likely that the drawn ball is contained inside, and hence the more likely that it will be given to the robot. In fact, the chance that a box is received by the robot is proportional to the number of balls contained inside.
- In the second step, with the knowledge of the color composition of the received box, the robot makes a guess about the color of the drawn ball following a fixed rule:

It guesses	Red	when the fraction of Red balls in the box received is	larger than X
	Blue		smaller than X
	Red/Blue		equal to X

The number X in the robot's rule is 50%-60% for the first 10 rounds, and 75%-85% for the last 10 rounds.

Monetary Payment

The monetary payment you receive for this experiment is determined as follows. The computer will randomly pick one round from the first 10 rounds, and another round from the latter 10 rounds. The sum of the payoff points that you received in these two randomly-selected rounds will be converted to money at an exchange rate of 1 point = 1 HKD, and paid to you. As every round may count, we recommend that you act in each round *as if* it is a round that counts.