ABSTRACT

Title of Dissertation: ESSAYS ON INFORMATION AND GENDER

Elif Bike Osun
Doctor of Philosophy, 2023

Dissertation Directed by: Professor Erkut Y. Ozbay
Department of Economics

Chapter 1 studies the effect of different feedback structures on belief updating in an ego-relevant task using a controlled experiment. Across treatments, subjects receive feedback through a signal with either a noise component, a comparison component, or both. The first two signals are commonly used in the literature, while I develop the latter to systematically analyze the effect of noise and comparison components on belief updating. I find that the signal structure is an important determinant of how subjects update their beliefs. This is driven by men and women exhibiting different biases depending on whether the signal is noisy or comparative. Men underweight bad news when the signal has a noise component and women underweight good news when the signal has a comparison component. These findings have implications for policies aiming to reduce the well-established gender gap in self-confidence through feedback provision.

In Chapter 2, I experimentally investigate whether there is a gender difference in advice giving in a gender-neutral task with varying difficulty in which the incentives of the sender and the receiver are perfectly aligned. I find that women are more reluctant to give advice compared

to men for difficult questions. The gender difference in advice giving cannot be explained by gender differences in performance. Self-confidence explains some of the gender gap, but not all. The gender gap disappears if advice becomes enforceable. I discuss possible underlying mechanisms that are consistent with the findings.

Voluntary disclosure literature suggests that in evidence games, where the informed sender chooses which pieces of evidence to disclose to the uninformed receiver who determines his payoff, commitment has no value, as there is a theoretical equivalence of the optimal mechanism and the game equilibrium outcomes. In Chapter 3, Erkut Ozbay and I experimentally investigate whether the optimal mechanism and the game equilibrium outcomes coincide in a simple evidence game. Contrary to the theoretical equivalence, our results indicate that outcomes diverge and that commitment has value. We also theoretically show that our experimental results are explained by accounting for lying averse agents.

ESSAYS ON INFORMATION AND GENDER

by

Elif Bike Osun

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2023

Advisory Committee:
Professor Erkut Y. Ozbay, Chair/Advisor
Professor Emel Filiz-Ozbay, Co-Advisor
Professor Yusufcan Masatlioglu
Professor Daniel R. Vincent
Professor Emanuel Zur

Dedication

I dedicate this dissertation to my father, Arif Osun, who was my biggest supporter in life, my role-model, and a constant source of love and encouragement.

Acknowledgments

I would like to express my gratitude to my main advisor, Erkut Ozbay, for his utmost support in every dimension of my academic and non-academic problems, for being an exemplar of how to be a good researcher, and for his endless generosity and kindness. I am forever indebted to him. I am also extremely grateful to my co-advisor, Emel Filiz-Ozbay, who has been an incredible source of inspiration and motivation, helping me navigate the challenges of graduate school and providing me with insightful feedback on my work.

I am very thankful to Yusufcan Masatlioglu for always keeping his door open and providing me with invaluable advice throughout graduate school. I would like to thank Daniel Vincent for many useful conversations and feedback on my research that he has provided. Furthermore, I am grateful to Emanuel Zur for generously giving his time and serving on my committee.

I thank all faculty, staff members, and students who have contributed to my research and overall experience at the University of Maryland. I am extremely grateful to Vickie Fletcher, for approaching every graduate student with kindness and always checking in on us. She is the rock of the Economics Department. I am thankful to John Shea for being an advocate for graduate students during difficult times and for his useful feedback on my various research projects. I am thankful to Neslihan Uler for her kindness and many helpful conversations. I would like to thank Guido Kuersteiner and Boragan Aruoba for their help and advice in navigating the job market.

I am grateful to classmates and members of the Behavioral and Experimental Economics

Brownbags group for creating a supportive research environment and for engaging in many helpful discussions. I am especially thankful to Prateik Dalmia and Ian Chadd for leading the way and always being open to giving me advice. I would also like to acknowledge the friendship and support I received from Keaton Ellis, Zhenxun Liu, Nathalie Gonzalez, and Macarena Kutscher.

I was lucky to have a support system outside of the University of Maryland as well. I am deeply appreciative of Larry Ausubel and Kathy Jacobson for offering me a wonderful job experience. I feel fortunate to have had the opportunity to work under their guidance. I am thankful to Yesim Orhun for her mentoring and all the time she spent with me discussing research ideas, and to Billur Aksoy for her wonderful recommendations and support over the years.

None of this would have been possible without the unending support of my family and friends. I am forever indebted to my parents, Arif and Selnur for their love, support, and the value they placed on my education. I am thankful to my stepmother Ozlem, for her care and kindness. I thank my two wonderful brothers, Ata and Arda, who have been sources of encouragement and goofiness even in difficult times. I thank my sister, Aysegul, for her unconditional love. I am extremely grateful for my wonderful friends - Irmak, Irmak, Gizem, and Doga - who have been a loving constant in my life for more years than I can count. I am very thankful to Ece, Semih, Olga, Ferda, Diego, and Kaan for becoming my family away from home. A special thanks to Chebu for being a very good boy.

Finally, I cannot express my gratitude enough to my husband, Berk. Thank you for being there for me through all the ups and downs, for embracing the countless foster animals that I bring into our home, for moving across continents with me three times, and for having been willing to do it again if needed. It's wonderful to know that we will face whatever life brings together, and I look forward to it.

Table of Contents

List of Tables

List of Figures

Chapter 1: Effect of Feedback on Beliefs About Self-Ability

1.1 Introduction

Beliefs about one's own ability shape important life decisions, but there is overwhelming evidence that individuals do not have accurate beliefs about themselves.[1] Distorted beliefs about one's ability can be costly, as they affect economically relevant choices such as which major to declare, which career path to choose, or salary negotiation upon a job offer. One way to correct for misaligned beliefs is to give feedback. However, predicting the effect of feedback on beliefs about ability is not straightforward. The theoretical benchmark for belief updating is Bayes' rule, yet experimental evidence from economics and psychology shows that individuals deviate from Bayes' rule in various ways. Studies focusing on belief updating in ego-relevant domains have not yet reached a consensus on the effect of receiving good versus bad news on how individuals update their beliefs.[2] Understanding the effect of feedback provision is valuable to make well-informed policy recommendations.

In a typical ego-relevant belief updating experiment, subjects complete a task, submit prior beliefs about their relative performance, receive some form of feedback on their performance, and submit posterior beliefs after observing their feedback. There are two signal structures commonly

[1] For example, 88% of U.S. drivers rate themselves safer than the median driver (Svenson, 1981) and only 3.8% of subjects in Niederle and Vesterlund (2007) guess they have the worst performance in a group of 4 people.

[2] I discuss studies documenting belief updating deviations from Bayes' rule in ego-relevant contexts in more detail in Section 1.2. For a survey of deviations from Bayesian updating in a broader context, see Benjamin (2019).

used in the literature for feedback provision: *noisy* and *comparative* signals.[3] Imagine a subject who took a test and is trying to guess whether they are among top or bottom half of performers among a group of individuals who completed the same task. A *noisy* signal reveals whether the subject is among top or bottom half with some accuracy rate, sometimes erroneously revealing the incorrect state. A *comparative* signal truly reveals whether the subject performed better or worse than a randomly chosen opponent among those who completed the same task. It is not clear whether the differences between the signal structures themselves affect updating behavior, yet the two signals are used interchangeably in ego-relevant belief updating experiments. In this paper, I systematically analyze the effect of these two commonly used signal structures on belief updating.

I design an experiment to compare belief updating under a *noisy* signal to that under a *comparative* signal structure. In the first part of the experiment, subjects complete an ego-relevant IQ task. I place the subjects in a group of other individuals who completed the same task and rank them based on their performance. In the second part of the experiment, subjects submit their beliefs on their relative performance twice: once before (prior beliefs) and once after they receive some feedback (posterior beliefs). I use a behavioral model based on Grether (1980), which uses weighted log-likelihood ratios of prior beliefs, good news, and bad news to construct posterior beliefs. Estimating the weight on each component allows me to detect deviations from the weights under the Bayesian benchmark.

I vary the signal structure used to generate feedback and compare belief updating behavior across treatments. A direct comparison between updating behavior under *noisy* and *comparative*

[3]For example, Buser et al. (2018), Coutts (2019), Schwardmann and Van der Weele (2019), Barron (2021), Möbius et al. (2022) use a *noisy* signal, while Eil and Rao (2011), Zimmermann (2020), Coffman et al. (2021b), Drobner (2022) use a *comparative* signal to provide feedback.

signals raises two issues. First, the two signals do not necessarily have the same informational content: the informativeness of the *noisy* signal is determined by the accuracy rate of the signal and is the same for all subjects, but the informativeness of the *comparative* signal varies by subject, as it depends on the subject's prior belief distribution over ranks. Second, the *noisy* signal has a noise component but lacks a comparison component, while the reverse is true for the *comparative* signal. Hence, there is a two-dimensional change across the two signals. To address these two issues, I design a novel signal structure that determines the accuracy of a noisy signal by comparing the subject to a randomly chosen opponent (so it has both noise and comparison components), in such a way that the informational content of the signal is isomorphic to that in the treatment with a *noisy* signal.[4] Implementing a signal that has both components allows me to detect the effect of noise and comparison in a controlled way by changing one component at a time.

Gender may affect how individuals update beliefs about their own ability. Gender gaps in labor market outcomes remain persistent, with women earning 83 cents on the dollar relative to men (Shrider et al., 2021) even though women make up more than half (50.7%) of the college-educated labor force in the United States (Fry, 2022). A large body of experimental literature documents robust gender differences in self-confidence, with men displaying more overconfidence than women (Barber and Odean, 2001; Niederle and Vesterlund, 2007). This gender difference in self-confidence may contribute to the well-established gender gap in labor market outcomes through human capital choices.[5] Feedback provision can be a valuable policy intervention to

[4]The behavioral model that I use incorporates the informativeness of the signals, hence a comparison between *noisy* and *comparative* signals is still possible despite the differences in information contents across treatments. By generating an additional signal that has a comparison component but is informationally isomorphic to the *noisy* signal, I am able to rule out information differences across treatments to be the driving mechanism of updating differences across treatments.

[5]For example, Cortés et al. (2021) find that gender differences in overconfidence lead to differences in job search

shrink the gender gap in labor market outcomes, yet what type of feedback is the most appropriate for this purpose is an open question. It is possible that men and women react differently to the two aforementioned feedback structures. Women are documented to dislike competition (Niederle and Vesterlund, 2007), so they might react to signals generated through comparison differently than men. Furthermore, men are shown to attribute bad news to luck while women attribute it to ability (Shastry et al., 2020), so getting a signal with a noise component might also have a differential effect on belief updating by gender.

Indeed, gender differences in belief updating have been documented in the literature. Coffman et al. (2019) explore how feedback affects gender differences in self-assessments and find that men and women exhibit different updating patterns upon receiving feedback, depending on the gender-congruency of the task. They document that men update their beliefs more optimistically than women if the task is male-typed (and vice versa if the task is female-typed). Since subjects have higher self-confidence for their performance on tasks that are in their gender's domain to begin with, receiving feedback in this setup actually fuels persistence in the gender gap in self-confidence. The focus of Coffman et al. (2019) is the gender-congruency of the task and not the type of the signal used to provide feedback. In this paper, I examine whether signals with noise and comparison components affect belief updating behavior differently for men and women, as this could help in designing policies to reduce the gender gap in self-confidence.[6]

I find that using different signal structures affects belief updating behavior. Although isomorphic in their informational content, receiving a signal with only a noise component leads to different deviations from the theoretical benchmark compared to receiving a signal with both

behavior of men and women college students.

[6]In Section 1.2, I discuss other papers that document gender differences in belief updating behavior.

noise and comparison components. The difference is driven by men and women exhibiting different updating behavior depending on whether the signal has a noise or comparison component. I examine updating behavior of men and women separately under three treatments. I find that women never underweight bad news and men never underweight good news. In contrast, how women update under good news and how men update under bad news is sensitive to signal type. Men underweight bad news in both treatments in which the signal has a noise component, whereas women underweight good news in both treatments in which the signal has a comparison component. These findings imply that for policies aiming to shrink the well documented gender gap in self-confidence, providing feedback with a noise component is not ideal if bad news is more prevalent, whereas providing feedback with a comparison component is not ideal if good news is more common. I conduct an ex-post analysis on gender differences in posterior beliefs and find suggestive evidence in line with these policy implications.

This paper contributes to the literature in several ways. This is the first study to systematically analyze the effect of different feedback structures on belief updating in a unified framework. I generate a novel signal structure that allows comparison between the effect of noise and comparative components of feedback in a controlled manner.[7] This is also the first paper documenting gender differences in how men and women perceive news under different signal structures. My findings suggest that policies aiming to reduce the gender gap in self-confidence by providing feedback on performance should carefully take the feedback structure and the performance of the target population into account; otherwise, providing feedback might be ineffective.[8]

[7] Coutts et al. (2020) independently developed a similar feedback structure to examine self-serving bias when updating beliefs under multiple sources of uncertainty.

[8] In fact, performance feedback may sometimes lead to worse outcomes. For example, Azmat et al. (2019) find that providing relative performance feedback decreases students' educational performance in a higher education setting field experiment.

The remainder of this paper is organized as follows. Section 1.2 discusses the related literature. Section 1.3 introduces the experimental design, treatments, and experimental protocol. Section 1.4 explains the methodology for measuring biases in belief updating and presents the results. Section 1.5 concludes.

1.2 Related Literature

The main assumption of the neoclassical theory of probabilistic beliefs is that upon receiving new information, individuals revise their beliefs according to Bayes' rule. Early experiments in the psychology literature documenting deviations from Bayes' rule using hypothetical belief updating questions include Edwards (1968) and Kahneman and Tversky (1972). These studies provide exogenous priors and compare the subjects' posterior beliefs to the Bayesian benchmark for the given signals about the underlying state. In contexts such as beliefs about self-ability, the prior beliefs are endogeneous and heterogeneous across subjects, so comparing posterior beliefs to the Bayesian benchmark is not sufficient to determine the source of updating deviations. Grether (1980) introduced a model of belief updating that allows one to detect deviations from the Bayesian weights on prior beliefs and signals separately through parameter estimation. A number of recent studies on belief updating, including this paper, use Grether's model to detect updating deviations from Bayes' rule (e.g. Möbius et al., 2022; Barron, 2021).

In the context of belief updating when information is ego-relevant, such as information about one's ability, the theoretical literature proposes several models to explain deviations from Bayes' rule. Landier (2000) proposes a model in which beliefs have a hedonic component through anticipation utility. Köszegi (2006)'s subjects derive ego utility from positive views about their

ability to do well in a skill-sensitive task. Mayraz (2009) provides an axiomatic model in which beliefs are affected by desires. More recently, Möbius et al. (2022) build a model of optimally-biased Bayesian updating. The common prediction of all these models is that good news is weighted more than bad news.[9]

Experimental studies focusing on belief updating in ego-relevant domains lack consensus on the weights assigned to good versus bad news when updating beliefs. Eil and Rao (2011) are among the first to document belief updating deviations for good versus bad news. They study updating in response to news about beauty and intelligence, and find that subjects give more weight to good compared to bad news. Möbius et al. (2022) and Drobner (2022) also find that positive information is weighted more heavily than negative when updating beliefs in an IQ-related quiz. In contrast, Ertac (2011) examines updating in response to news about performance on tasks requiring ability and effort and finds that individuals incorporate bad news more into their beliefs than good news. Coutts (2019) finds that bad news receives more weight compared to good news when updating beliefs in ego-relevant, financially-relevant, and neutral domains. Grossman and Owens (2012) document that subjects have overconfident beliefs about their performance on an intelligence-based task, but their belief updating follows the Bayesian benchmark upon receiving both good and bad news. Barron (2021) uses a financially-relevant task that is not ego-relevant but with payoffs such that subjects prefer one state over the other, and also finds updating in line with Bayes' rule. Buser et al. (2018), Schwardmann and Van der Weele (2019), and Zimmermann (2020) find that subjects do not give enough weight to their signals when updating their beliefs

[9]Confirmation bias is another mechanism proposed to explain deviations from Bayes' rule. Rabin and Schrag (1999) build a model of confirmation bias, in which individuals give more weight to information that conforms with their prior beliefs. However, the experimental studies (including this paper) did not find direct evidence for confirmation affecting belief updating in ego-relevant domains. For example, Eil and Rao (2011) find that valence is the underlying cause of confirmatory bias and that confirmation alone has no effect. Möbius et al. (2022) examine and find no evidence of confirmation bias.

relative to Bayes' rule, but give equal weight to good and bad news.[10] All the studies mentioned here use a single feedback structure in their experimental design, and none examine the effect of the feedback structure on belief updating.

A few papers propose mechanisms that can contribute to the lack of consensus in belief updating behavior across experimental studies. Drobner (2022) shows that expectations about resolution of uncertainty affect belief updating behavior. Using an IQ test and exogeneously manipulating subjects' expectations about the resolution of uncertainty, he finds that those who are informed that their true rank will not be revealed at the end of the experiment update their beliefs optimistically, while those who are informed that they will learn their true rank at the end of the experiment update their beliefs neutrally. Coffman et al. (2019) examine subjects' beliefs about their performance on tasks that vary in their gender-congruency and document that gender stereotypes influence belief updating: subjects give more weight to good news over bad when the signal arrives in a gender-congruent domain. Coutts (2019) uses a feedback structure that noisily informs subjects whether they were among the top 15% of performers and finds that subjects give more weight to bad news compared to good news when updating their beliefs. Even though his experiment is not designed to test the effect of the informational content of signals on belief updating, he considers the use of negatively skewed signals as an ex-post explanation of bad news receiving more weight than good news. In this paper, I consider the type of signal structure used to give feedback as another mechanism that can affect belief updating behavior.

In addition to Coffman et al. (2019), discussed in the introduction, several papers document gender differences in belief updating. Ertac (2011) finds that women update their beliefs more

[10] Zimmermann (2020) also examines belief updating in the long run. The results reported here are the findings immediately after feedback. In the long run, he finds that the effect of receiving good news persists, but the effect of bad news fade over time.

pessimistically, by giving less weight to good news compared to men. The gender difference in belief updating arises only in the verbal GRE task, which was perceived as more difficult by the subjects, but not in the easy algebraic addition task. Möbius et al. (2022) and Coutts (2019) find that women update their beliefs more conservatively than men both for good and bad news, but are not significantly more asymmetric. In Coutts et al. (2020), men are significantly more responsive to good news relative to bad news when they receive feedback about their own performance, while women do not update their beliefs asymmetrically. Coffman et al. (2021b) examine the effect of feedback on beliefs in a dynamic setting and show that while both men and women underweight both type of signals, the effect of bad news on men's beliefs fades more over time compared to women's beliefs, leading to persistent gender differences in self-confidence in the long run. Similar to other papers studying belief updating, these studies use a single feedback structure in their experimental design and are not designed to test the gender differences in belief updating across signal structures.

Signals with noise and comparative components are used interchangeably in the literature.[11] This paper is concerned with investigating the effect of different feedback structures on belief updating and their differential effect by gender. One cannot address this question by sorting the existing literature on the type of the signal used and making a meta-analysis, as various other aspects of the experimental design differ across studies, including type of the task (e.g. SAT questions, logic questions, raven's matrices, ASVAB questions, summation task, beauty task) and the performance measure used for belief elicitations (e.g. the likelihood of being among top or bottom performers, the likelihood of being among a pre-determined percentile, expected

[11] Grossman and Owens (2012), Buser et al. (2018), Coutts (2019), Schwardmann and Van der Weele (2019), Barron (2021), Möbius et al. (2022) use a noisy signal. Ertac (2011) uses a variation of the comparative signal in which the comparison is not against a single opponent, but against a group of opponents. Eil and Rao (2011), Zimmermann (2020), Coffman et al. (2021b), Drobner (2022) use a comparative signal.

rank, absolute score). Hence, there is need for a controlled experiment to investigate the effect of different signal structures on belief updating in a unified framework.

1.3 Experimental Design

I designed the experiment using the experimental software oTree (Chen et al., 2016) and conducted it online on Prolific during April and May 2022. I recruited 901 subjects from the U.S. subject pool.[12] No subject participated in the experiment more than once. Median completion time was about 10 minutes and median payment was about $13 per hour excluding the completion fee.[13]

The experiment consisted of four parts, detailed below, and an exit questionnaire. Figure 1.1 summarizes the timeline of the experiment. In the first part of the experiment, subjects completed an IQ task. Upon completing the test, subjects were informed that they were randomly placed in a group of 9 other participants who previously solved the same test.[14] Then, subjects submitted their beliefs on their relative performance among this group, both before and after they received feedback. The experiment had a between-subject design, with each treatment using a different signal structure for feedback provision.

[12]From this initial pool, I drop the data of 9 subjects whose reply to the survey question about their gender is inconsistent with their demographic data on Prolific, one subject who revoked their consent after completing the study, and one subject who timed out and as a result could not be paid.

[13]The completion fee was $1.1 for about one third of the participants and was increased to $1.4 for the remaining two third after Prolific increased the minimum hourly participant reward from $6.5 to $8 on April 21, 2022. The total payment including the bonus payments was already above the updated minimum required hourly payment, however the platform makes the minimum payment calculation at the time of announcing the study, and does not take the bonus payments into account. Rather than changing the duration of the experiment to keep the completion fee the same, I increased the completion fee and kept the duration of the experiment the same after the price change.

[14]The 9 other participants for each subject were randomly chosen from a group of Prolific participants who previously completed the same IQ test before data collection for the main experiment began.

Figure 1.1: Timeline of the experiment

1.3.1 Test Stage

Subjects had four minutes to answer as many questions as possible. The test consisted of questions typically used to measure IQ, an ego-relevant belief domain. Questions were standard logic questions similar to those used in Möbius et al. (2022) and Cognitive Reflection Test questions (Frederick, 2005), such as:

1. Which one of the five choices makes the best comparison? LIVED to DEVIL as 6323 is to:

a) 2336 b) 6232 c) 3236 d) 3326 e) 6332

2. If some Wicks are Slicks, and some Slicks are Snicks, then some Wicks are definitely Snicks. The statement is: a) True b) False c) Neither

3. In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? a) 7 days b) 13 days c) 24 days d) 47 days e) 48 days

Earnings for the quiz were $0.20 per question answered correctly, so the subjects had a monetary incentive to perform as well as possible. At the test stage, subjects knew that there was going to be another part of the experiment, but they did not know the content of the following part. Hence, subjects did not know that they would submit their beliefs about their relative performance compared to a group of other participants when solving the quiz. This was to avoid any incentive

to perform poorly in order to guarantee having correct beliefs about relative performance later. Subjects did not learn their earnings until the end of the experiment, so they could not make any inferences about their performance from their quiz earnings.

1.3.2 Prior Belief Elicitation Stage

Once the test stage was over, subjects were informed that their performance would be compared to 9 other randomly chosen participants who previously completed the same quiz. To examine belief updating, I focus on subjects' beliefs about whether they were among the top half or bottom half of performers in their group, which is a subjective probability over a binary type space, as updating a single number is more intuitive compared to updating a distribution. However, I also measure subjects' belief distribution over ranks, to calculate the likelihood ratio of the signals when the feedback structure is comparative.[15] For these purposes, subjects were asked the following questions:

Question 1. How do you estimate the likelihood (in percent) of being in each rank when your performance is compared to the other 9 members of your group?

Question 2. What do you think is the likelihood (in percent) that you rank among the top and bottom halves of the performers in the group? In other words, in the group of 10, what do you think is the likelihood that your rank is 1, 2, 3, 4, or 5 (you are among the top half performers) and what do you think is the likelihood that your rank is 6, 7, 8, 9, or 10 (you are among the bottom half performers)?

I asked the above two questions on the same screen to eliminate concerns about anchoring. The experimental interface was split in two parts. The left side of the screen consisted of the first

[15]The signal types and calculation of likelihood ratios are explained in more detail in Subsection 1.3.3

question and allowed subjects to submit their beliefs for the likelihood of being in each rank. The right side of the screen consisted of the second question, and the probabilities of being among the top and bottom half of performers were calculated in real time as subjects modified their answer to the first question. Once subjects submitted their beliefs, they were asked to confirm that the likelihood of being among the top and bottom half performers reflected their true beliefs in a separate screen, to further increase the salience of the second question. Subjects could go back to the previous screen to edit their answer if they wished (see Appendix A.3 for screenshots).

To eliminate hedging motives, either prior or posterior beliefs were randomly chosen for payment. I incentivized prior beliefs using the quadratic scoring rule (Selten, 1998) with the following formula:

$$100 - 50 \times \sum_{i=1}^{10} (\mathbb{1}[rank = i] - \frac{p_i}{100})^2 \tag{1.1}$$

where $\mathbb{1}[rank = i]$ is an indicator variable that takes the value 1 if subject's rank was equal to i and 0 otherwise, and p_i is their estimate for being in rank $i \in 1, 2, ..., 10$.

Note that subjects were incentivized for their estimates on the likelihood of each rank and not separately for their likelihood of being among the top or bottom half, as incentive compatibility in one question leads to incentive compatibility in the other. An alternative would be to incentivize both questions on prior beliefs separately and randomly choose one to implement. I avoided this in order to minimize the complexity of information I provided to subjects on the screen.

Even though incentive compatibility of the quadratic scoring rule requires assuming risk neutrality, there are several reasons to suggest that this is not an obstacle for interpeting the results of this paper. First, possible earnings from each belief elicitation question ranged from $\$0$ to $\$1$,

stakes over which one would expect risk-neutrality. Secondly, similar to Eil and Rao (2011) and other papers using the quadratic scoring rule (e.g. Zimmermann, 2020; Barron, 2021), I explicitly told the subjects that they would maximize their expected earnings if they report their true beliefs. Thirdly, Danz et al. (2020) show that truthful likelihood reporting is maximized when subjects are not provided with the exact formula for the payoff calculation. Following this argument, the main experimental screen did not include the explicit formula for payoff calculation, but only included the sentence "Your expected payoff will be the highest if you report your true beliefs." The interested subjects could click on a link to access the exact formula. Lastly, the main results that I focus on in this paper compare belief updating behavior across signal structures. Any tendency to hedge beliefs due to risk preferences in one treatment would likely be the same in other treatments, having no effect on the relative bias across treatments.

1.3.3 Feedback Stage

The signal structure used to provide feedback varied by treatment and had either a noise component, a comparison component, or both. I call these treatments *Noisy*, *Comparative*, and *NoisyComparative*, respectively. In all of the treatments, subjects received instructions about how their signal would be determined and needed to answer a comprehension question correctly before receiving feedback.

1.3.3.1 Noisy Treatment

Feedback in the *Noisy* Treatment consisted of a signal with an accuracy rate of 7/9: if a subject was among the top half of performers of their group, they would receive a signal stating

that they were among the top half performers (good news) with probability 7/9, and a message stating that they were among the bottom half performers (bad news) with probability 2/9. If a subject was among the bottom half of performers of their group, they would receive bad news with probability 7/9 and good news with probability 2/9. In this treatment, the meaning of the signal has a "noise" component in the sense that it is incorrect with some probability. It does not have a "comparison" component, since the signal is not determined through the subject being compared to another individual. This signal structure is commonly used in the belief updating literature (e.g. Buser et al. (2018), Coutts (2019), Barron (2021), and Möbius et al. (2022)).

Table 1.1: Signals in Noisy Treatment

Performance	**Signal Received**
Top half	"Top half" with 7/9 chance "Bottom half" with 2/9 chance
Bottom half	"Bottom half" with 7/9 chance "Top half" with 2/9 chance

1.3.3.2 Comparative Treatment

The signals in the *Comparative* Treatment informed subjects whether their performance was better (good news) or worse (bad news) than a randomly chosen participant in their group. Hence, the signal has a comparison component. There is no noise component in the meaning of the signal, as it always conveys correct information. This is another signal structure commonly used in the belief updating literature (e.g. Eil and Rao (2011), Zimmermann (2020), Coffman et al. (2021b), and Drobner (2022)).

Table 1.2: Signals in Comparative Treatment

Performance	Signal Received
Better than randomly chosen participant	"Better than the other participant"
Worse than randomly chosen participant	"Worse than the other participant"

Comparison Between Noisy and Comparative Treatments: There are two issues with directly comparing belief updating behavior between the *Noisy* and *Comparative* treatments. First, the informativeness of the signals under the two treatments are not the same. The likelihood ratios of receiving good and bad news in the *Noisy* Treatment are homogeneous across subjects and are determined by the accuracy rate of the signal:

$$LR_G^N = \frac{Pr(s = G \mid t = H)}{Pr(s = G \mid t = L)} = \frac{7}{2} \quad , \quad LR_B^N = \frac{Pr(s = B \mid t = H)}{Pr(s = B \mid t = L)} = \frac{2}{7} \tag{1.2}$$

where LR_G^N and LR_B^N are the likelihood ratios of receiving a good and a bad signal in the *Noisy* Treatment, s is the signal received (where G and B denote good and bad news), and t is the performance type (where H and L denote being among the top half and bottom half of performers).

The likelihood ratios of receiving good and bad news in *Comparative* Treatment, in contrast, are determined by the subject's prior beliefs over ranks, and hence vary by subject. Denote p_i as the prior probability given to being in each rank $i = \{1, 2, ..., 10\}$ (where 1 is the best performance and 10 is the worst performance). The probability of randomly choosing a participant with worse performance (i.e. receiving a good signal) given rank i in a group of 9 others is $(10 - i)/9$, and the probability of randomly choosing a participant with better performance (i.e. receiving a

bad signal) is $(i-1)/9$. Then,

$$Pr(s=G|t=H) = \frac{Pr((s=G)\cap(t=H))}{Pr(t=H)} = \frac{p_1\times\frac{9}{9}+p_2\times\frac{8}{9}+p_3\times\frac{7}{9}+p_4\times\frac{6}{9}+p_5\times\frac{5}{9}}{Pr(t=H)} \tag{1.3}$$

$$Pr(s=G|t=L) = \frac{Pr((s=G)\cap(t=L))}{Pr(t=L)} = \frac{p_6\times\frac{4}{9}+p_7\times\frac{3}{9}+p_8\times\frac{2}{9}+p_9\times\frac{1}{9}+p_{10}\times\frac{0}{9}}{Pr(t=L)} \tag{1.4}$$

$$Pr(s=B|t=H) = \frac{Pr((s=B)\cap(t=H))}{Pr(t=H)} = \frac{p_1\times\frac{0}{9}+p_2\times\frac{1}{9}+p_3\times\frac{2}{9}+p_4\times\frac{3}{9}+p_5\times\frac{4}{9}}{Pr(t=H)} \tag{1.5}$$

$$Pr(s=B|t=L) = \frac{Pr((s=B)\cap(t=L))}{Pr(t=L)} = \frac{p_6\times\frac{5}{9}+p_7\times\frac{6}{9}+p_8\times\frac{7}{9}+p_9\times\frac{8}{9}+p_{10}\times\frac{9}{9}}{Pr(t=L)} \tag{1.6}$$

Using the above equations, the likelihood ratios of receiving good and bad news in the *Comparative* Treatment are:

$$LR_G^C = \frac{Pr(s=G|t=H)}{Pr(s=G|t=L)} = \frac{p_1\times\frac{9}{9}+p_2\times\frac{8}{9}+p_3\times\frac{7}{9}+p_4\times\frac{6}{9}+p_5\times\frac{5}{9}}{p_6\times\frac{4}{9}+p_7\times\frac{3}{9}+p_8\times\frac{2}{9}+p_9\times\frac{1}{9}+p_{10}\times\frac{0}{9}} \times \frac{Pr(t=L)}{Pr(t=H)} \tag{1.7}$$

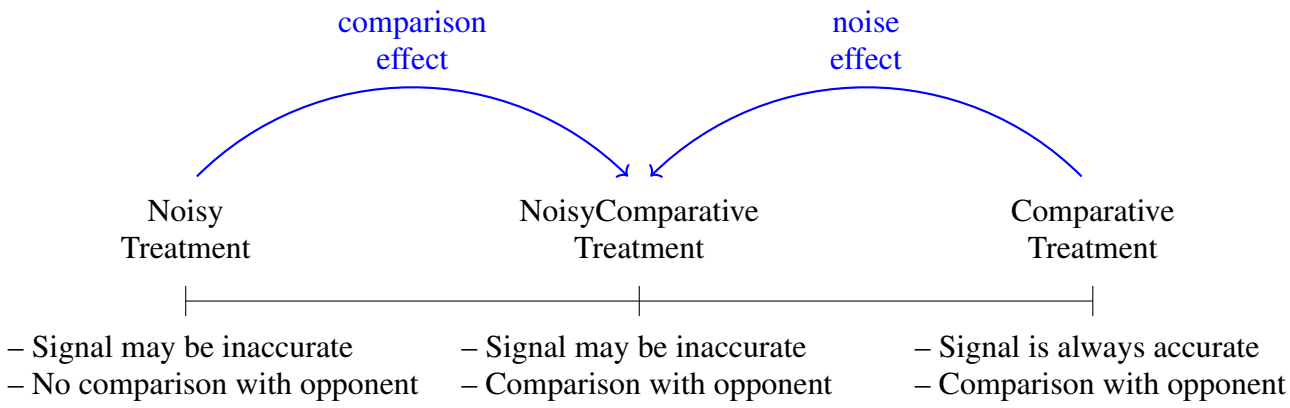
$$LR_B^C = \frac{Pr(s=B|t=H)}{Pr(s=B|t=L)} = \frac{p_1\times\frac{0}{9}+p_2\times\frac{1}{9}+p_3\times\frac{2}{9}+p_4\times\frac{3}{9}+p_5\times\frac{4}{9}}{p_6\times\frac{5}{9}+p_7\times\frac{6}{9}+p_8\times\frac{7}{9}+p_9\times\frac{8}{9}+p_{10}\times\frac{9}{9}} \times \frac{Pr(t=L)}{Pr(t=H)} \tag{1.8}$$

Since it is possible that $LR_G^N \neq LR_G^C$ and $LR_B^N \neq LR_B^C$, as can be seen by comparing equations (1.2), (1.7), and (1.8), the informativeness of the signals are not necessarily the same under the *Noisy* and *Comparative* Treatments. I design a novel signal structure that determines the accuracy rate of a noisy signal by comparing the subject to a randomly chosen opponent in such a

way that the informational content of the signal is isomorphic to that in the *Noisy* Treatment, as I explain in further detail in the following subsection. The behavioral model that I use incorporates the informativeness of the signals when examining belief updating behavior, so a comparison between *Noisy* and *Comparative* Treatments is still possible, yet having a signal structure with a comparison component that is also informationally isomorphic to the signal in the *Noisy* Tratment allows me to rule out information differences across treatments being the driving mechanism of updating differences across treatments.

The second issue with directly comparing belief updating behavior between the *Noisy* and *Comparative* Treatments is that there are two changes across treatments: the signal in the *Noisy* Treatment has a noise component but lacks a comparison component, while the the reverse is true for the signal in the *Comparative* Treatment. As illustrated in Figure 1.2, the *NoisyComparative* Treatment acts as a bridge between the two treatments, allowing me to consider the effect of one change at a time.

Figure 1.2: Summary of treatments



1.3.3.3 NoisyComparative Treatment

In the *NoisyComparative* Treatment, the accuracy of the signal is determined by comparing the subject's performance type to a randomly chosen participant's. If the subject and the randomly chosen opponent are in different halves of the distribution (i.e. if one is among the top half while the other is among the bottom half of performers), then the signal correctly reveals whether the subject is among the top or bottom. If the subject and the randomly chosen opponent are in the same half of the distribution (i.e. if both are among the top half or both are among the bottom half of performers), then the signal has a 50% chance of revealing the correct type and 50% chance of revealing the incorrect type.

Table 1.3: Signals in NoisyComparative Treatment

Subject Performance	**Opponent Performance**	**Signal Received**
Top half	Bottom half	"Top half"
Bottom half	Top half	"Bottom half"
Top half	Top half	"Top half" with 50% chance "Bottom half" with 50% chance
Bottom half	Bottom half	"Top half" with 50% chance "Bottom half" with 50% chance

To see that the informational content of the signals in the *NoisyComparative* Treatment is equivalent to those in the *Noisy* treatment, note that:

$$\begin{aligned} Pr(s = G \mid t_i = H) &= Pr(t_{-j} = L \mid t_j = H) \times 1 + Pr(t_{-j} = H \mid t_j = H) \times 1/2 \\ &= 5/9 \times 1 + 4/9 \times 1/2 = 7/9 \end{aligned} \tag{1.9}$$

where s is the signal received (where G and B denote good and bad news), t is the performance type (where H and L denote being among the top half and the bottom half of performers), j is the index for the subject, and $-j$ is the index for the randomly chosen opponent. Similarly,

$$Pr(s = G \mid t_j = L) = Pr(t_{-j} = H \mid t_j = L) \times 0 + Pr(t_{-j} = L \mid t_j = L) \times 1/2 = 2/9 \quad (1.10)$$

$$Pr(s = B \mid t_j = H) = Pr(t_{-j} = L \mid t_j = H) \times 0 + Pr(t_{-j} = H \mid t_j = H) \times 1/2 = 2/9 \quad (1.11)$$

$$Pr(s = B \mid t_j = L) = Pr(t_{-j} = L \mid t_j = L) \times 1/2 + Pr(t_{-j} = H \mid t_j = L) \times 1 = 7/9 \quad (1.12)$$

Hence, the likelihood ratios of receiving good and bad news in the *NoisyComparative* Treatment are equivalent to those calculated in Equation (1.2):

$$LR_G^{NC} = \frac{Pr(s = G \mid t = H)}{Pr(s = G \mid t = L)} = \frac{7}{2} \quad , \quad LR_B^{NC} = \frac{Pr(s = B \mid t = H)}{Pr(s = B \mid t = L)} = \frac{2}{7} \quad (1.13)$$

1.3.4 Posterior Belief Elicitation Stage

I elicited subjects' posterior beliefs about the likelihood of being among the top and bottom half of performers of their group after they observe their signal. I do not elicit beliefs over ranks in this stage, as I only need beliefs over being among the top half or the bottom half of performers to examine subjects' updating behavior.

As discussed above, either prior or posterior beliefs were randomly chosen for payment to prevent hedging motives. I incentivized posterior beliefs using the quadratic scoring rule (Selten,

1998) with the following formula:

$$100 - 50 \times \sum_{k \in \{top, bottom\}} (\mathbb{1}[half = k] - \frac{p_k}{100})^2 \tag{1.14}$$

where $\mathbb{1}[half = k]$ is an indicator variable that takes the value 1 if the subject was among the k half of performers in the group and 0 otherwise, and p_k is the subject's posterior likelihood of being among half $k \in \{top, bottom\}$.

1.4 Results

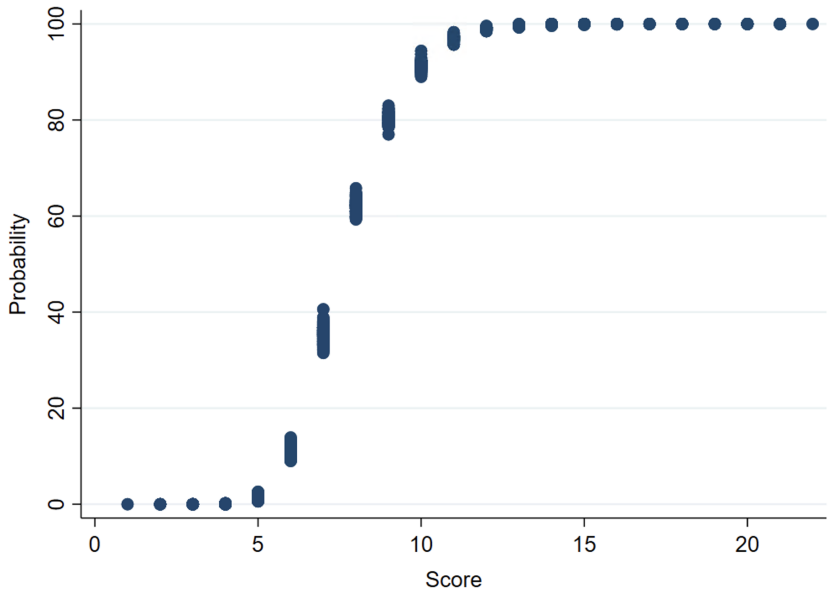
From the main data, I exclude some observations to minimize noise stemming from lack of comprehension or not paying attention to instructions. Similar to previous studies (e.g. Möbius et al., 2022, Barron, 2021, Coutts, 2019), I exclude subjects who reported posterior beliefs that were updated in the opposite direction compared to the Bayesian prediction (i.e. an upward shift in the belief of being among top performers after a bad signal or a downward shift in the belief of being among top performers after a good signal). These observations correspond to 9.9% of the subjects, which is in line with findings from previous studies. Secondly, given the online nature of the experiment, subjects who did not read the instructions require caution. I exclude subjects who spent less than 10 seconds both on the screen with instructions about the signal structure and on the comprehension question screen (in which instructions about the signal structure were also accessible), resulting in the exclusion of 2.4% of the remaining subjects.[16]

[16] I choose the 10 seconds cutoff in an ad-hoc manner, aiming for a lower bar on how fast the signal structure summary page can be read with comprehension. The main results are qualitatively similar if no subject is excluded based on time spent on instructions.

1.4.1 Overview of Prior Beliefs and Confidence

As a preliminary analysis, I examine subjects' prior beliefs relative to their actual performance. As a belief accuracy benchmark, I generate each subject's bootstrapped probability of being among the top half of performers given their score. I run a simulation with 1,000 repetitions in which I randomly match each subject with 9 other participants and generate an indicator variable representing whether the subject was among the top half of performers. The bootstrapped probability of subject k being among the top half is $\sum_{r=1}^{1000} d_{r,k}/1000$, where $d_{r,k}$ is an indicator variable that takes the value 1 if subject k was among the top half performers of their group in the r^{th} replication of the simulation, and 0 otherwise. Figure 1.3 depicts the bootstrapped probability for each score. The bootstrapped probability is a cleaner benchmark for subjects' belief accuracy compared to simply using a dummy variable indicating whether they were among the top half of performers in their experimental group, as two subjects with the same score can have different realized outcomes due to luck.

Figure 1.3: Bootstrapped probability of being among the top half of performers given score



The mean absolute error in prior beliefs (calculated by taking the absolute value of the difference between the subject's prior belief and the bootstrapped probability of their being among the top half of performers) is equal to 35.5 points and is significantly different than 0 ($p < 0.001$) using a Wilcoxon signed-rank test.[17] In line with previous findings, I find that subjects do not have accurate beliefs about their relative performance on average.

I also find a significant gender difference in self-confidence, even though men and women perform similarly on the IQ test. On average, men and women answer 8.46 and 8.25 questions correctly, respectively. The difference is not statistically significant ($p = 0.427$). Figure 1.4 illustrates men and women's prior beliefs for each score, illustrating that women have lower priors for each possible score. Table 1.4 provides further evidence that women have significantly lower self-confidence, based on OLS regressions relating prior beliefs to gender and actual performance

[17]Unless otherwise stated, all p-values to compare distributions are obtained using the Mann Whitney U-test, while all p-values to compare measures to benchmarks are obtained using the Wilcoxon signed-rank test.

$(p < 0.001)$.

Figure 1.4: Prior beliefs of being among the top half of performers by actual score

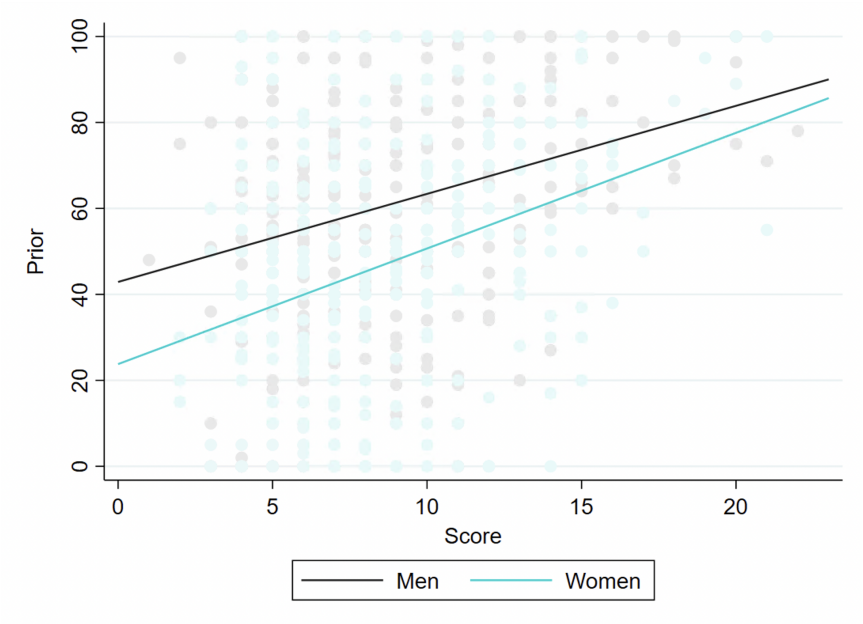


Table 1.4: OLS regressions relating prior beliefs to gender and performance

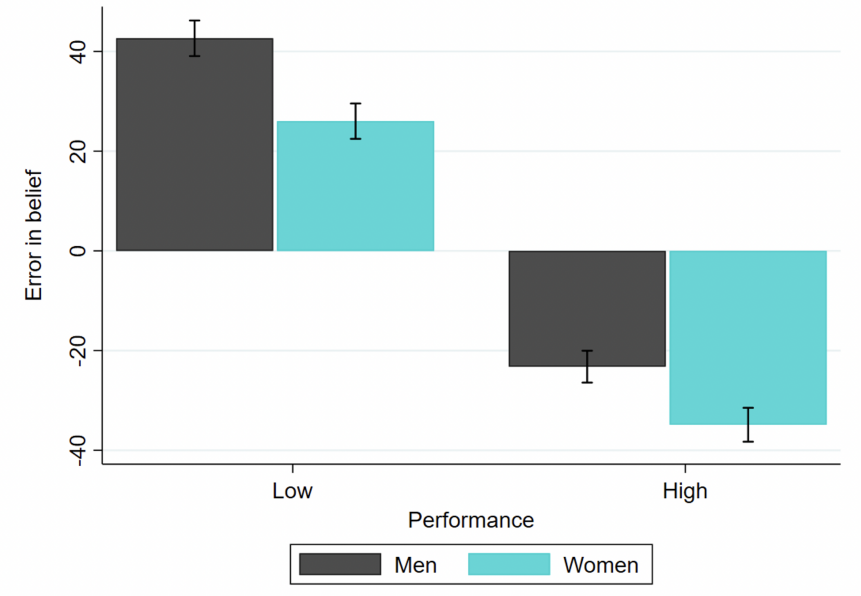
Prior	(1)	(2)	(3)
Female	-14.2***	-13.8***	-19.1***
	(2.020)	(1.923)	(4.737)
Score		2.4***	2.0***
		(0.259)	(0.357)
Female*Score			0.641
			(0.519)
Constant	60.2***	40.3***	42.9***
	(1.411)	(2.569)	(3.301)
N	783	783	783

Notes: Prior is the prior belief of being among the top half of performers. *Female* is a dummy variable equal to 1 if gender is female and 0 if male. *Score* is the number of questions answered correctly. *Female*Score* is the interaction of gender and score. Standard errors are in parentheses, * p<0.1, ** p<0.05, *** p<0.01.

Finally, I find significant gender differences in over and underconfidence. The mean error in prior beliefs, calculated as the difference between the prior belief and the bootstrapped

probability of being among the top half of performers, is 9.1 points for men and -2.8 points for women. Note that positive values correspond to overconfidence and negative values correspond to underconfidence. The difference across genders is significant $(p < 0.001)$. Since low performers are more likely to get bad news and high performers are more likely to get good news, I also examine the gender difference in self-confidence by performance level. I classify low performers as those with less than a 50% probability of being among the top half and high performers as those with more than a 50% probability of being among the top half of performers. As Figure 1.5 illustrates, both men and women with low performance are overconfident, but men are more overconfident than women (the mean errors in prior beliefs are 42.6 vs 26.0, $p < 0.001$). Both men and women with high performance are underconfident, but women are more underconfident than men (the mean errors in prior beliefs are -23.2 vs -34.9, $p < 0.001$).

Figure 1.5: Mean error in prior beliefs by gender and performance



1.4.2 Belief Updating Behavior

I investigate belief updating behavior using the model originated by Grether (1980), which maintains the general Bayesian structure but allows for different weights on the prior, good news, or bad news compared to the Bayesian benchmark.[18] Consider the following likelihood ratio:

$$\frac{Pr(H|S)}{Pr(L|S)} = \left(\frac{Pr(H)}{Pr(L)}\right)^{\delta} \times \left(\frac{Pr(S|H)}{Pr(S|L)}\right)^{\beta} \tag{1.15}$$

where H *(High)* and L *(Low)* correspond to being among top and bottom half of performers among the reference group, $Pr(H|S)$ and $Pr(L|S)$ are posterior beliefs given signal $S \in \{Good, Bad\}$, and $Pr(H)$ and $Pr(L)$ are prior beliefs of being among top half and bottom half of performers.[19] In the standard Bayesian model, $\delta = \beta = 1$. Adding indicator variables to distinguish between good and bad news, Equation 1.15 becomes:

$$\frac{Pr(H|S)}{Pr(L|S)} = \left(\frac{Pr(H)}{Pr(L)}\right)^{\delta} \times \left(\frac{Pr(S=G|H)}{Pr(S=G|L)}\right)^{\beta_G \times \mathbb{1}[S=G]} \times \left(\frac{Pr(S=B|H)}{Pr(S=B|L)}\right)^{\beta_B \times \mathbb{1}[S=B]} \tag{1.16}$$

where G and B denote receiving a signal with good news and bad news, respectively. Finally, log-linearizing Equation 1.16 allows me to test for behavioral biases on priors and signals using an OLS regression:

$$ln\,(posterior) = \delta \times ln\,(prior) + \beta_G \times \mathbb{1}[S=G] \times ln\,(LR_G) + \beta_B \times \mathbb{1}[S=B] \times ln\,(LR_B) \tag{1.17}$$

[18] This model is also used by Möbius et al. (2022), Coutts (2019), Coffman et al. (2019), and Holt and Smith (2009).

[19] In all notation, I use type (H)igh and (L)ow to represent being among top and bottom halves instead of (T)op and (B)ottom. This is to avoid confusion with the notation of (G)ood and (B)ad signals.

where $ln\,(posterior) = ln\,(Pr(H|S)/Pr(L|S))$ is the posterior log-likelihood ratio for being among the top half given signal $S \in \{Good, Bad\}$. A positive value indicates allocating higher probability to being among the top half and a negative value indicates allocating higher probability to being among the bottom half. δ is the weight given to prior log-likelihood ratio for being among the top half, $ln\,(prior) = ln\,(Pr(H)/Pr(L))$. LR_G and LR_B are the likelihood ratios of observing good and bad news, respectively.[20] $\mathbb{1}[S = G]$ and $\mathbb{1}[S = B]$ are indicator variables that are equal to 1 for the corresponding signal and 0 otherwise. β_G and β_B measure the responsiveness of the posterior to receiving good and bad news, respectively. Borrowing from the nice summary provided by Benjamin (2019) and Barron (2021) on interpreting the values of the δ, β_G, and β_B coefficients, Table 1.5 presents various belief updating biases documented in the literature.

Table 1.5: Interpretation of OLS Coefficients

Coefficient	Interpretation
$\delta = \beta_G = \beta_B = 1$	Bayesian updating
$\delta < 1$	Base-rate neglect
$\delta > 1$	Base-rate overuse
$\beta_G < 1$ or $\beta_B < 1$	Conservatism
$\beta_G > 1$ or $\beta_B > 1$	Overinference
$\beta_G \neq \beta_B$	Asymmetry

This behavioral model is silent with regard to prior and posterior beliefs at the boundary (i.e. beliefs equal to 0% or 100%). Following Charness and Dave (2017), Holt and Smith (2009) and Grether (1992), I truncate the data so that beliefs about being among the top and bottom performers lie in the interval [1%, 99%]. I replace posterior beliefs equal to 0% with 1% and

[20] I show the calculations of LR_G and LR_B for each treatment in Subsection 1.3.3. See Equations (1.2), (1.7), (1.8), and (1.13).

those equal to 100% with 99%. For prior beliefs over ranks, I replace probabilities of 0% with 0.2% and subtract the total added probability from all non-zero probability ranks, weighted by the prior in the corresponding rank.[21]

$$p_i^* = \begin{cases} 0.2 & \text{if } p_i = 0 \\ p_i - \frac{p_i \times 0.2 \times n_0}{100} & \text{if } p_i \neq 0 \end{cases} \tag{1.18}$$

where p_i is the prior belief on rank $i \in \{1, 2, ..., 10\}$ before truncation, p_i^* is the same belief after truncation, and n_0 is the number of ranks with 0 prior belief. Truncated prior beliefs over being among the top half and bottom half of performers are the sum of the truncated prior beliefs over relevant ranks ($\sum_{i=1}^{5} p_i^*$ for top, $\sum_{i=6}^{10} p_i^*$ for bottom).[22]

1.4.2.1 Belief Updating Compared to the Bayesian Benchmark

The neoclassical theory of probabilistic beliefs predicts that all of the coefficients of Equation (1.17) are equal to 1, corresponding to Bayes' rule. Behavioral models such as ego utility or confirmation bias predict deviations from Bayes' rule, yet there is no existing theoretical model that predicts differential updating behavior across treatments based on the signal structure. Using the analysis described above, I compare updating behavior to the Bayesian benchmark across treatments. Table 1.6 reports the coefficients from estimating the OLS regression in Equation (1.17). The upper part of the table reports coefficients and their corresponding standard errors. A coefficient significantly different than 0 (as indicated by stars) indicates that prior beliefs and

[21] The truncation of beliefs over being among the top half and bottom half of performers prevents the $ln(posterior)$ and $ln(prior)$ terms (in all treatments) and the truncation of beliefs over ranks prevents the LR_G and LR_B terms (in the *Comparative* Treatment) from blowing up in Equation (1.17).

[22] I consider other truncation methods as well, which are explained in Appendix A.2. All results are robust to using the alternative truncation methods.

receiving good or bad news significantly affect posterior beliefs. As expected, all coefficients δ, β_G, and β_B are significantly different than 0 ($p < 0.001$). The bottom half of Table 1.6 compares estimated coefficients to the Bayesian benchmark. Any coefficient different than 1 is a deviation from Bayes' rule.

Table 1.6: Belief updating across treatments

Regressor	N (1)	NC (2)	C (3)
δ	0.666***	0.731***	0.718***
	(0.030)	(0.037)	(0.028)
β_G	0.937***	0.770***	0.725***
	(0.075)	(0.081)	(0.076)
β_B	0.878***	0.703***	0.815***
	(0.077)	(0.089)	(0.094)
p-values for H_0 :			
$\delta = 1$	0.000	0.000	0.000
$\beta_G = 1$	0.377	0.008	0.000
$\beta_B = 1$	0.126	0.001	0.094
$\beta_G = \beta_B$	0.580	0.570	0.499
N	261	255	267
R^2	0.752	0.718	0.745

Notes: Columns (1)-(3) report results from the OLS regression on *Noisy* Treatment, *Noisy-Comparative* Treatment, and *Comparative* Treatment, respectively. The first half of the table reports coefficient values and their associated standard errors below in parentheses with * p<0.1, ** p<0.05, *** p<0.01. The second half of the table reports p-values from Chow-tests on equality of coefficients to 1 or to each other.

Subjects exhibit base-rate neglect in all treatments ($p < 0.001$ for $H_0 : \delta = 1$) and do not update asymmetrically in any of the treatments ($p > 0.1$ for $H_0 : \beta_G = \beta_B$). The existence of conservatism varies by signal structure. There is no evidence of conservatism in the *Noisy* Treatment ($p = 0.377$ for $H_0 : \beta_G = 1$, $p = 0.126$ for $H_0 : \beta_B = 1$), while there is conservative updating of both good and bad news in the *NoisyComparative* Treatment ($p = 0.008$ for $H_0 : \beta_G = 1$, $p = 0.001$ for $H_0 : \beta_B = 1$), and conservative updating of only good news in the

Comparative Treatment ($p < 0.001$ for $H_0 : \beta_G = 1$, $p = 0.094$ for $H_0 : \beta_B = 1$) at the 5% level. Even though the *Noisy* and *NoisyComparative* treatments are informationally isomorphic, adding a comparison component to the noisy signal results in different updating behavior across treatments.

Result 1 *Updating behavior is sensitive to the signal structure, even when the informational content of the two signals is equivalent. Subjects do not update conservatively in the Noisy Treatment but exhibit conservatism in the NoisyComparative and Comparative Treatments.*

1.4.2.2 Belief Updating Across Genders

Next, I examine whether there is a gender difference in how noisy and comparative signals are processed. Table 1.7 reports the results from estimating the OLS regressions based on Equation (1.17) separately for each gender and signal structure. The upper part of the table reports coefficients and their corresponding standard errors. Again, all coefficients are significantly different than 0, verifying that prior beliefs, good news, and bad news significantly affect posterior belief formation for both genders in all treatments. The bottom half of Table 1.7 reports coefficients compared to the Bayesian benchmark for men and women in each treatment. Any coefficient different than 1 is a deviation from Bayes' rule.

Table 1.7: Belief updating across treatments by gender

	N		NC		C	
Regressor	Men (1)	Women (2)	Men (3)	Women (4)	Men (5)	Women (6)
δ	0.715***	0.589***	0.714***	0.725***	0.718***	0.703***
	(0.044)	(0.040)	(0.063)	(0.047)	(0.044)	(0.038)
β_G	0.919***	0.921***	0.814***	0.737***	0.778***	0.654***
	(0.104)	(0.106)	(0.128)	(0.104)	(0.116)	(0.102)
β_B	0.673***	1.107***	0.629***	0.775***	0.758***	0.908***
	(0.118)	(0.100)	(0.141)	(0.120)	(0.136)	(0.132)
p-values for H_0 :						
$\delta = 1$	0.000	0.000	0.004	0.000	0.000	0.000
$\beta_G = 1$	0.456	0.403	0.203	0.023	0.063	0.000
$\beta_B = 1$	0.004	0.320	0.004	0.053	0.119	0.544
$\beta_G = \beta_B$	0.139	0.194	0.366	0.803	0.922	0.172
N	133	128	133	122	135	132
R^2	0.745	0.783	0.663	0.777	0.729	0.771

Notes: Columns (1)-(2), (3)-(4), and (5)-(6) report results from the OLS regression using the data of men and women separately from *Noisy* Treatment, *NoisyComparative* Treatment, and *Comparative* Treatment, respectively. The first half of the table reports coefficient values and their associated standard errors below in parentheses with * p<0.1, ** p<0.05, *** p<0.01. The second half of the table reports p-values from Chow-tests on equality of coefficients to 1 or to each other.

I find that how women update their beliefs upon receiving good news and how men update their beliefs upon receiving bad news is sensitive to signal type. Men underweight bad news in both treatments in which the signal has a noise component ($p = 0.004$, 0.004, 0.119 for $H_0 : \beta_B = 1$ in *Noisy*, *NoisyComparative*, and *Comparative* Treatments, respectively), whereas women underweight good news in both treatments in which the signal has a comparison component ($p = 0.403$, 0.023, 0.000 for $H_0 : \beta_G = 1$ in *Noisy*, *NoisyComparative*, and *Comparative* Treatments). Furthermore, men do not significantly underweight good news in any treatment at the 5% level, ($p = 0.456$, 0.203, 0.063 for $H_0 : \beta_G = 1$ in *Noisy*, *NoisyComparative*, and *Comparative* Treatments), while women do not underweight bad news in any treatment ($p = 0.320$, 0.053, 0.544 for $H_0 : \beta_B = 1$ in *Noisy*, *NoisyComparative*, and *Comparative* Treatments).

Result 2 *Noise and comparison components in a signal have a differential effect on belief updating by gender. Men underweight bad news if the signal has a noise component, whereas women underweight good news if the signal has a comparison component. Regardless of the signal structure, women do not underweight bad news and men do not underweight good news.*

1.4.2.3 Policy Implications

The findings on belief updating differences across genders indicate that for policies aiming to reduce the gender gap in self-confidence, providing feedback with a noise component may not be ideal in environments in which bad news is more prevalent, since men underweight bad news when the signal has a noise component, while women do not. Similarly, providing feedback with a comparison component may not be ideal in environments in which good news is more prevalent, since women underweight good news when the signal has a comparison component, while men do not. I test these conjectures by examining the gender difference in posterior beliefs across treatments separately for recipients of good and bad news. I run the following OLS regression of posterior beliefs on gender, priors, test score, and other individual characteristics:

$$posterior_j = \beta_0 + \beta_F \times female_j + \beta_P \times prior_j + \beta_S \times score_j + \gamma \times C_j + \epsilon_j \qquad (1.19)$$

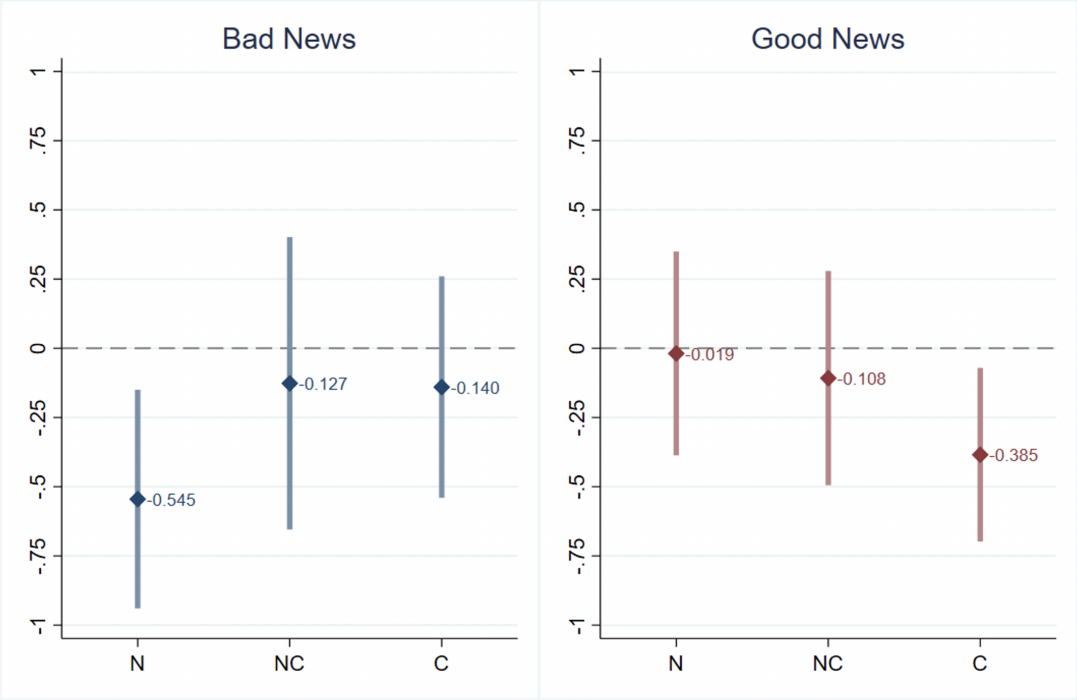
where $posterior$ is the posterior log-likelihood ratio for being among the top half of performers, $female$ is a dummy variable equal to 1 if the gender is female and 0 if male, $prior$ is the prior log-likelihood ratio for being among the top half, $score$ is the number of correct answers in the IQ test, C is a vector of individual characteristics including age, education, and income, and j

denotes the subject index.[23]

Figure 1.6 plots the β_F coefficient in Equation (1.19), which captures the gender difference in posterior beliefs after feedback provision. In line with the conjectures given above, the largest gender gap in posterior beliefs for those who receive bad news is in the *Noisy* Treatment, while the largest gender gap in posterior beliefs for those who receive good news is in the *Comparative* Treatment. Women who receive bad news in the *Noisy* Treatment and women who receive good news in the *Comparative* Treatment have significantly lower posterior beliefs compared to men after controlling for score and prior beliefs (with $p\text{-}values$ 0.007 and 0.017, respectively). The complete list of coefficients and their corresponding standard errors are depicted in Table A.3.

[23]The number of men and women who receive good news is balanced across treatments. Testing the equivalence of percent of men and women who receive good news with a test of proportions yields p-values 0.157, 0.987, and 0.572 for *Noisy*, *NoisyComparative*, and *Comparative* Treatments, respectively. See Table A.2 for the breakdown of number of subjects.

Figure 1.6: Gender Gap in Posterior Beliefs Across Treatments



Notes: N, NC, and C correspond to *Noisy* Treatment, *NoisyComparative* Treatment, and *Comparative* Treatment, respectively. The figure illustrates β_F coefficient of Equation (1.19) with 95% confidence intervals.

The findings indicate that receiving bad news in the *Noisy* Treatment and good news in the *Comparative* Treatment result in a gender gap even after controlling for prior beliefs. Given that women have lower prior beliefs on being among the top half of performers compared to men, the documented gender differences on posterior beliefs can be seen as a lower bound on the adverse effects of receiving noisy bad news or comparative good news. If prior beliefs are dropped from the set of controls, the gender differences in the *Noisy* Treatment under bad news and in the *Comparative* Treatment under good news become even more pronounced. Furthermore, the gender gap in posteriors for both types of news in the *NoisyComparative* Treatment becomes significant at the 5% level, also in line with predictions based on the belief updating patterns

documented in Subsection 1.4.2.2. However, when the data is broken down by treatment, gender, and signal type, we are left with smaller sample sizes and there are some imbalances in prior beliefs across subgroups. Hence, I focus on the regressions conditioning on prior beliefs in the main body of the paper. The regression results not conditioning on priors can be found in Appendix Table A.4.

1.5 Conclusion

People are not great at forming accurate beliefs about their abilities, which leads to sub-optimal economically-relevant decisions. Giving performance feedback is one way to correct for misaligned beliefs, but there is no consensus on how individuals update their beliefs. In this paper, I show that the structure of feedback is an important factor affecting belief updating biases. The results of my controlled experiment show that the weights subjects give to good and bad news vary by whether the signal has a noise component or a comparison component. In previous work examining belief updating biases, noisy and comparative signals have been used interchangeably.

A gender breakdown of misaligned beliefs shows that men have higher self-confidence than women with similar abilities, a result commonly found in previous studies. Furthermore, men and women react differently to signals with a noise or a comparison component. I find evidence that men underweight bad news when receiving noisy signals, while women underweight good news when receiving comparative signals. Understanding how feedback mechanisms affect gender differences in belief updating can help us design more efficient policies to shrink the gender gap in self-confidence through feedback provision.

This paper shows that the feedback structure affects updating behavior in a controlled ex-

perimental setting. It is still an open question whether one would observe the same effects when belief updating is tied to making choices that could affect earnings in higher stakes. Investigating whether these belief distortions translate into actions, such as selecting into competition, or into real life decisions using a field experiment are promising directions for future work.

Chapter 2: Gender Differences in Advice Giving

2.1 Introduction

There is a well-documented but not fully explained gap between the labor market outcomes of men and women. Women are underrepresented in higher management positions. In S&P 500 companies where 44.7% of all employees are women, the percentage of women steadily decreases for higher-profile positions, with the female percentage of CEOs at 5.8% (Catalyst, 2020). Management positions usually require giving advice to employees, peers, and supervisors. If women are reluctant to give advice compared to men, this could contribute to having fewer women in higher-profile positions, either through self-selection or by making women less qualified for the job. Exploring whether such a gap exists is of interest.

There are several factors that might play a role in women being underrepresented in higher management positions. Discrimination is one potential contributing factor that has been explored in the literature both theoretically and experimentally (see for e.g., Lazear and Rosen, 1990; Bohren et al., 2019; Coffman et al., 2021c). Aside from discrimination, underlying preference differences may also be responsible for having fewer women in higher-profile positions. For example, women are shown to be less competitive than men (Niederle and Vesterlund, 2007), less willing to act as the decision maker of a group in stereotypically male-typed domains (Coffman, 2014), volunteer more for low-promotability tasks (Babcock et al., 2017), and do not self-promote

as much as men (Exley and Kessler, 2019), all of which can contribute to the gender gap in career advancement. This paper: (i) explores whether women are more reluctant to give advice compared to men as another potential underlying preference difference contributing to the gender gap, (ii) investigates the effect of task difficulty on the gender difference to give advice, and (iii) investigates the effect of enforceability of advice on decision to send it. The first point aims to identify whether such a gender gap exists, and the latter two aims to provide insight to the mechanisms contributing to the gender gap in advice giving.

Using a gender-neutral task with varying difficulty, I show that female senders are less likely to send their guess as advice compared to men, but only when the question is difficult. The results indicate that women are not always reluctant to give advice, but shy away from giving advice when the question is difficult. I find that performance and self-confidence both have a significant effect on advice giving, but the gender discrepancy cannot be fully explained even after controlling for factors such as performance, self-confidence, risk preferences, and demographics. Finally, I show that there is no significant gender gap in a group decision making setup in which the advice becomes enforceable. In light of these results, I discuss possible underlying mechanisms of the documented gender gap.

The first objective of this study is to investigate whether there is an inherent difference in willingness to give advice between men and women absent potential confounds. This could best be achieved using a controlled experiment. Even though one can imagine ways make the experimental design more complex to mimic a real-life manager position (such as accounting for familiarity of the subjects, allowing for backlash, using male-typed or female-typed tasks), this is not the aim of this paper. There is a trade-off between the complexity of the experiment and potential confounds on the variable of interest. My aim is to explore whether there is a gender gap

in preference to give advice, motivated by advice giving being an important part of managerial positions. Hence, I use a simple experimental design to focus on investigating such a gap in a controlled environment.

In addition to identifying a gender gap in willingness to send advice, the experimental variation in difficulty of the task and enforceability of advice provides additional insights on mechanisms contributing to the gender difference in advice giving. To my knowledge, this is the first paper systematically analysing the relationship between the difficulty of a task and the propensity to give advice. Findings suggest that women are not simply averse to sharing their opinion, but they are unwilling to do so in situations which they are less sure about how correct their advice is. As a possible underlying mechanism, I consider if women disliking to decide on behalf of others (as in Ertac and Gurdal, 2012, Erat and Gneezy, 2012) is the driver of the gender difference in advice giving. I test and rule out this explanation by varying the enforceability of advice and showing that there is no significant gender gap in advice giving when advice becomes enforceable, rather than the gender gap increasing. This is also the first paper to show that the gender differential in advice giving is affected by whether advice is enforceable. Even though the treatments in this paper cannot identify the exact mechanism leading to gender differences in advice giving, I develop a model of guilt and responsibility which can explain the findings of the experiment and can serve as a guide for further investigation to identify the underlying mechanisms in future work.

The remainder of this paper is organized as follows. Section 2.2 discusses the related literature. Section 2.3 presents the experimental design and procedures. Section 2.4 explores the nature of the task. Section 2.5 presents the experimental results. Section 2.6 introduces a simple model of guilt and responsibility. Section 2.7 conducts robustness analysis. Section 2.8

concludes.

2.2 Related Literature

Group decision making and leadership literatures are closely related to advice giving and there are several studies exploring gender differences in these contexts. Ertac and Gurdal (2012) show that women are less likely to volunteer as the group leader to decide on a risky investment task on behalf of their group. Coffman (2014) examines the subjects' willingness to act as the decision maker of a group for tasks that are stereotypically outside of their gender's domain. The Dictator Treatment in this paper is closely related to and complements her main finding. I find that there is no gender difference to act as the decision maker of a group when the task is gender-neutral, while Coffman (2014) uses male-typed and female-typed tasks and finds that both men and women are less willing to act as the decision maker of a group when the task is outside of their gender's domain. Considering that women are typically documented to be more-risk averse than men, if the investment task in Ertac and Gurdal (2012) is perceived as a male-typed task, this may be driving the result of women's unwillingness to be the leader in their environment. Other studies show that introducing backlash disproportionately deters women from self-selecting into leadership roles (Chakraborty et al., 2021), the gender composition of a group affects the subjects' willingness to be the leader (Born et al., 2020), this effect of the gender composition is mainly through the salience of gender stereotype of the task (Chen and Houser, 2019), and gender incongruency of a task plays a significant role for choosing the decision maker of a group (Coffman et al., 2021a). The gender gap in advice giving that I document in this paper indicates that for managerial positions which require advice giving, gender differences may arise

even for tasks that are not gender-incongruent.

The advice literature generally focuses on the receiving end of the advice. The experimental findings suggest that subjects have a tendency to follow advice, want to receive advice when given the option, and make better decisions in the presence of advice –even when the advisor is not an expert for the task at hand (see Schotter, 2003 for a survey on the effect of naive advice on decision making). Advice has been shown to increase cooperation (Chaudhuri et al., 2009), improve learning (Iyengar and Schotter, 2008; Çelen et al., 2010), influence selection into competition (Brandts et al., 2015), help subjects with strategic play (Cooper and Kagel, 2016), and affect truthful revelation in school matching mechanisms (Ding and Schotter, 2017; 2019). The effect of advisor's gender on advice seeking has also been studied. There is some evidence of gender-based discrimination on the value of advice. Nyarko et al. (2006) show that female advisors suffer a discount in a market where clients compensate advisors. Yet, Heikensten and Isaksson (2019) find that the advisor's gender does not affect the advisee's willingness to seek costly advice. Manian and Sheth (2021) investigate how advice from different advisors is perceived and find that even though advice is not discriminated against based on gender, subjects expect women advisors to be followed less than men. These papers focus on the effect of advice on decision makers or how advice is perceived, whereas I focus on understanding the preferences for advice giving.

Compared to advice receiving, there are fewer studies which explore advice giving behavior. Gneezy et al. (2020) investigate advice giving when advisors' incentives are biased towards one of the two investments they can recommend to a client. The focus of their paper is the bias in advice giving rather than the preference to send advice, so advisors in their design do not have the option not to send advice. In a large-scale field experiment on high school students,

Eskreis-Winkler et al. (2019) find that treated students who give motivational advice to younger students earn higher grades compared to those in control. Their main question is the effect of advice giving on advisors, so their design also does not have an option to not send advice for those assigned to treatment condition. Hinnosaar (2019) finds that women are less likely to contribute to Wikipedia than men, which may be related to women's unwillingness to give advice compared to men.

Cooper and Kagel (2016) examine why teams beat the benchmark of each group member's highest individual performance in signaling games. While their main focus is neither advice giving behavior nor gender differences, their findings are closely relevant to this paper. They find that advisor-advisee pairs do not perform as well as teams, which is driven by advisees not listening to sound advice as well as advisors not sending sufficient advice. The latter is driven mostly by female advisors. Even though the gender gap in advice giving observed in this signaling game is intriguing, strategic play may be affected by various factors such as beliefs about opponent's action. If men and women have different expectations about the action that their opponent will play in a strategic game, gender difference in these beliefs might systematically affect their decision to send advice. My paper differs from Cooper and Kagel (2016) in several ways. Firstly, I use a single person decision-making problem and abstract away from strategic considerations. Secondly, I experimentally vary the difficulty of the task and find that the gender difference in advice giving is limited to difficult questions. Finally, I show that the gender difference in the decision to send one's answer is affected by whether the receiver has to follow it. These results can shed additional light onto the mechanisms contributing to the gender gap in advice giving.

Most closely related to this paper, Brandts and Rott (2021) examine the effect of gender and gender matching on advice giving and advice following about entry into a real-effort tournament.

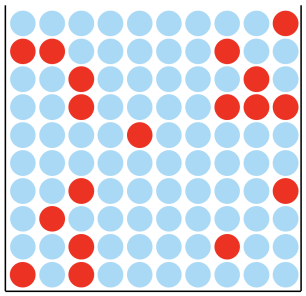
The advisors in their experiment choose whether to advise subjects to select into competing in a tournament or not. Even though the type of the task (math addition task versus ball counting task), the domain of advice (whether the advisee should enter a competition or not versus guessing the correct answer for the pair), and whether gender of the matched subject is explicitly mentioned varies across the two studies, the main findings are in line with each other. Brandts and Rott (2021) find that women are less likely than men to advise entering into competition; but only when the entrant has intermediate performance. This finding suggests that gender differences in advice giving emerge in situations that are more ambiguous. In line with this finding, I document that women are less likely than men to send advice, but only in difficult questions, whose correct answer, by design, should be harder to guess.

2.3 Experimental Design and Procedures

I conducted the experiment online via Amazon's Mechanical Turk (MTurk) between February 23 and April 4, 2021. I recruited 450 subjects from the U.S. subject pool. I used the experimental software oTree (Chen et al., 2016). No subject participated in the experiment more than once and the experiment had a between-subject design. The sessions lasted about 14 minutes on average. The average payment was $3.97 including the $1 completion fee. The experiment consisted of two parts and a survey. Appendix B.2 contains the instructions provided to the subjects. After the subjects saw the instructions of the first part, they had to answer three comprehension questions correctly to continue the experiment.[1] There were 25 rounds in the first part of the experiment, in which the task was to count the number of red balls in a box with 100 red and blue balls as depicted in Figure 2.1.

[1] See Figure B.13 for the exact wording of comprehension questions.

Figure 2.1: Box containing a mix of 100 red and blue balls



There were 5 easy, 10 medium, and 10 difficult questions. All subjects saw the questions in the same randomly generated order.[2] I classify questions as easy, medium, or difficult based on the number of red balls in the box. Table 2.1 depicts the difficulty levels of the questions based on the contents of the box. I expect the task to be more difficult as the numbers of red and blue balls get closer to each other.[3] I determined the cutoffs for each difficulty via a pilot conducted on graduate students at the University of Maryland during Experimental Economics Brownbags. The cutoffs were determined with the expectation that for most of the subjects, it would be possible to know the number of red balls at a glance in easy questions; it would be necessary to count the balls to know the correct answer in medium questions; and it would not be possible to count the exact number of red balls within 10 seconds in difficult questions. While the exact cutoff for a question to be classified as easy, medium, and difficult may vary by subject, I show that the average normalized errors are in line with my categorization of difficulty in Section 2.4 and I report additional robustness analyses around the cutoffs in Section 2.7.

[2]The exact questions and their order can be found in Table B.1.

[3]For example, I consider a box that contains 95 red balls as easy because the subjects can easily count the number of blue balls (5) and reach the correct answer (95) by subtracting the number of blue balls from 100.

Table 2.1: Difficulty of a question based on the number of red balls in the box

Difficulty	Number of red balls	Number of questions
Easy	$[0, 10)$ or $(90, 100]$	5
Medium	$[10, 30]$ or $[70, 90]$	10
Difficult	$(30, 70)$	10

Each image stayed on the screen for 10 seconds, after which the subjects were asked to submit their guess for the number of red balls in the box that they saw. To eliminate the concern that subjects could take a photo of the box and count the number of red balls without a time limit, the screens in which subjects were required to submit an answer were also limited to 10 seconds. To disincentivize subjects from leaving their screens unattended, they could not continue the experiment if they failed to submit an answer in 3 or more rounds due to timeout.[4]

The subjects were assigned one of two roles: *sender* or *receiver*.[5] The receivers were randomly matched with a new sender in each round to avoid reputation building. For each session, senders' data was collected first and asynchronously matched to receivers who completed the experiment later. I ran asynchronous sessions to overcome the challenges with subject dropouts frequently observed in online experiments (Zhou and Fishbach, 2016). Since I randomly match subjects in each round, running the sessions synchronously would require interacting a large number of subjects, which would have been a challenge with subject dropouts. Senders and receivers were matched using imperfect stranger matching. On average, each session consisted of 22 senders or receivers who were matched over a course of 25 periods, hence re-matching with the same subject occurred rarely.[6] Once a session with senders ended, senders' answers

[4]This occurred rarely; 0.5% of the subjects were dismissed from the experiment due to timing out more than twice.

[5]The terminology used in the experimental interface was "advisor" and "decision maker".

[6]Minimum number of subjects in a session was 17, maximum number of subjects in a session was 25.

were linked to the corresponding receivers' session. For each receiver in each round, one sender was randomly chosen as the match in that round. At the end of the experiment, the code randomly chose one round for payment for each sender-receiver pair in a session, with the constraint that resulting pairs would constitute a one-to-one matching (so that there would be unique sender-receiver pairs and each subject was matched to exactly one other person for the round that counts for payment). This constraint was never violated. At the time of making decisions, the only information available to the subjects about the person they would be matched with was that in each round, they would be randomly matched to another subject who was assigned the other role. Subjects were informed that the payment calculations would be done once all subjects in their session completed the experiment and that they would receive payment within 48 hours of completion.

Payoffs in the experiment were in terms of points with a conversion rate of 100 points = \$1. The payoff in each round was determined at the pair level (consisting of the sender and the receiver); the sender and the receiver in the same pair earned the same payoff. The payoff depended on the error in that round, which is the distance between the pair's final answer and the correct number of red balls in the box. The payoff was 400 points if the answer was correct, 200 points if the error was between 1-3, 100 points if the error was between 4-10, 50 points if the error was between 11-15, and 0 if the error was greater than 15. The gradually decreasing payoff structure (rather than an all-or-nothing payoff structure) aims to incentivize subjects to pay attention to the task even if they think that they cannot know the correct answer, which becomes more relevant as the question difficulty increases. This kind of payoff structure is common in real-life situations such as in employee bonuses or exam scores. At the end of the experiment, one round out of 25 was randomly selected for payment. Subjects did not receive any feedback

about their payoffs between rounds.

2.3.1 Treatments

There were two treatments in the experiment: *Advice Treatment* and *Dictator Treatment*, which differed in how the pair's final answer was determined. The subjects were balanced across treatments, role, and gender. There were either 55 or 56 subjects for all gender (male/female), role (sender/receiver), and treatment (Advice Treatment/Dictator Treatment) combinations.[7,8] The sequence of decisions in each treatment are summarized below (see Figures B.8-B.11 for the instructions provided to the subjects).

Advice Treatment: In each round, the senders saw the box on their screen for 10 seconds. They then submitted their guess for the number of red balls in the box. In the following screen, the senders chose whether to send their guess to the receiver or not. In each round, to incentivize senders to submit their guess truthfully, there was a 5% probability that the sender's guess was implemented as the pair's final answer. With 95% probability, the final answer of the pair was the receiver's guess. The receivers saw the same box that the senders saw for each round. If the sender sent their guess, the receiver saw the sender's guess before submitting theirs. Otherwise, the receiver was informed that the sender did not send their guess this round.

Dictator Treatment: The senders in this treatment also saw the box on their screen for 10 seconds, after which they submitted their guess for the number of red balls in the box. Then, they decided whether to send their guess to the receiver or not. As in the Advice Treatment, there was a 5% probability that the sender's guess was implemented as the pair's final answer. With

[7]The gender of the subjects were not revealed to one another throughout the experiment.

[8]There were 4 non-binary subjects whose data is excluded from the analysis. The breakdown of the number of subjects by gender, treatment, and role can be found in Table B.2.

95% probability, the sender's choice determined whose guess would be implemented as the pair's final answer. If the sender sent (didn't send) their guess, then the sender's (receiver's) guess was implemented as the pair's final answer. The receiver saw the same box that the senders saw for each round and then submitted their guess without any feedback about the sender's action. They learned whether the sender sent their guess to be implemented as the pair's final answer in the following screen.

Note that in the Advice Treatment, if the sender sends their guess, it is up to the receiver whether to follow the advice or not. The senders may have subjective beliefs about whether their advice will be followed or about how responsible they are for the pair's earnings if they send their guess in the Advice Treatment.[9] In order to make the sender's effect on the pair's earnings more pronounced and to have an insight on the underlying mechanisms of advice giving differences, I minimize the role of these subjective beliefs in the Dictator Treatment, in which the sender's guess is implemented as the final answer of the pair if they send it. Hence, the senders are effectively choosing whether to act as the pair's decision maker in the Dictator Treatment.

2.3.2 Other Tasks

At the end of the first part, I elicited the subjects' self-confidence using an incentivized rank guess as in Niederle and Vesterlund (2007), in which I ask the subjects where they thought they would rank in terms of submitting the most accurate answer among a group with 3 other randomly selected subjects of the same role for a randomly selected round. In the second part of the experiment, I elicited subjects' risk preference using an incentivized investment task following

[9]The senders did not receive any feedback about the receivers' answers nor about the extent their advice was followed.

Gneezy and Potters (1997), in which the subjects chose how much of their endowment to allocate between a safe and a risky option (see Figure B.14 and Figure B.15 for the screenshots of these two tasks).

2.4 Data

Before presenting the main experimental results, I first verify that the data is consistent with my assumptions on the difficulty of the questions and the gender-neutrality of the task. To measure subjects' performance, I use the normalized error, defined as the actual error divided by maximum possible error for a given question.[10] Unless otherwise stated, all p-values to compare distributions are obtained using the Mann Whitney U-test and all p-values to compare measures to benchmarks are obtained using the Wilcoxon signed-rank test throughout the paper. For all non-parametric tests, I compare measures generated at the individual level, so intercorrelation of observations of the same participant is not an issue.[11,12] For all regressions, I cluster standard errors at the individual level.

Recall that I aim to choose a gender neutral task so that any advice giving difference between genders is not an artifact of the nature of the task. The average normalized error of female (male) subjects is 11.7% (11.1%) for difficult questions, 4.8% (5.0%) for medium questions, and 1.6% (0.8%) for easy questions. The gender differences in performance are not significant for

[10] For example, if there were 5 red balls in the box and the subject guessed 7, then the normalized error would be $|7-5|/max\{5-0, 100-5\} \approx 2.1\%$. The results on both gender-neutrality of the task and categorization of difficulties are robust to using simply the absolute difference between the subject's guess and the correct answer instead of the normalized error.

[11] For example, to compare advice giving behavior in difficult questions using a non-parametric test, I first calculate the percent of difficult questions that the sender sent their guess at the individual level, and then test the equality of this ratio by gender.

[12] See Table B.2, which shows the number of men and women in each role and treatment, for the relevant number of observations used in a non-parametric test.

any difficulty level ($p > 0.1$ for all levels), supporting that inherent performance differences in gender are not likely to drive the differences in subjects' preference to give advice.

Next, I check whether my classification of task difficulty is appropriate. Recall that I set certain cutoffs to define the difficulty of tasks (as can be seen in Table 2.1). The average normalized errors of the subjects are 1.2%, 4.9%, and 11.4% for easy, medium, and difficult questions, respectively. The difference between each pair (easy-medium, medium-difficult, easy-difficult) is statistically significant ($p < 0.01$ for each pair). This indicates that the cutoffs that I use for classifying questions based on difficulty are appropriate.

For the results I report above, I use the performance of all senders in both treatments and of receivers only in the Dictator Treatment, since these subjects submit their guesses without receiving any external information beforehand.[13] Note that the receivers in the Advice Treatment submit their guesses after observing the sender's guess when available; hence, their performance may be affected by whether they received advice in a given round. I analyze the effect of advice on decision making separately in Section 2.5.4.

2.5 Experimental Results

I begin by analyzing the senders' decision to send their guess in both treatments. Section 2.5.1 explores whether women senders are less likely to send advice compared to men in the Advice Treatment and whether the gender gap can be explained by self-confidence and other demographic characteristics. Section 2.5.2 explores whether the gender gap persists in the Dictator Treatment, in which the senders' guess is enforced if they choose to send it. Section 2.5.3

[13]The results in this section use the combined data of these three groups of subjects. The total number of observations is 335, with 167 women and 168 men. The results are similar when each group is analyzed separately.

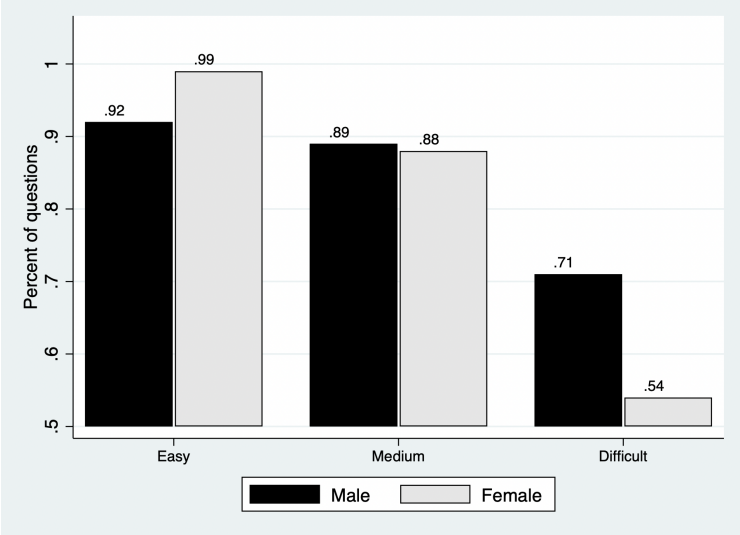
compares the behavior across treatments. Finally, I examine the effect of advice on receivers' performance, both in the aggregate data and when the data is broken down by gender in Section 2.5.4.

2.5.1 Do Women Shy Away From Giving Advice?

In order to determine whether women shy away from giving advice, I compare frequency of advice sending by male and female senders in the Advice Treatment for all questions and separately for each level of difficulty. Pooling all questions together, men send advice in 83% of the questions, while women do so in 77% of the questions. The difference is marginally significant at the 10% level ($p = 0.068$). When questions are broken down by difficulty level, different patterns emerge based on the difficulty of questions. Figure 2.2 shows the percentages of easy, medium, and difficult questions for which the senders sent their guess, broken down by gender. The gender difference in advice sending is not statistically significant for easy ($p = 0.113$) and medium ($p = 0.601$) questions, yet there is a significant gender gap for difficult questions ($p = 0.014$): women send their advice less frequently than men. On average, male senders send their guess for 71% of difficult questions, compared to only 54% for female senders. Figure B.1 plots the cumulative distribution functions (CDFs) of advice sending percentages on difficult questions for men and women senders, and shows that there is a first order stochastic dominance relationship between the distributions of men and women senders. Based on Somers' D statistic (Newson et al., 2001), a randomly chosen male sender is 27% more likely to send advice for a difficult question than a female sender, with a 95% confidence interval of $[0.047, 0.490]$, supporting the first order stochastic dominance relationship.

Further evidence can be found in Table 2.2, which reports the results of probit regressions relating advice sending to gender for a sample containing only difficult questions, clustering standard errors at the individual level. The regression shows that men are significantly more likely than women to send their guess as advice. The result is robust when I control for individual performance (as measured by the sender's normalized error on that question), self-confidence (as measured by their self-perceived rank of guess accuracy), demographics, and risk preferences.

Figure 2.2: Percent of Questions for which the Senders Send Their Guess, Advice Treatment



Notes: Figure illustrates percentages of easy, medium, and difficult questions for which the senders send their guess, broken down by gender. The p-values for the differences of percentages between men and women are $p = 0.113$ for easy, $p = 0.601$ for medium, and $p = 0.014$ for difficult questions.

Table 2.2: Probit Regressions Relating Advisor's Guess-Sending to Gender in Advice Treatment for Difficult Questions

Guess Sent	(1)	(2)
Male	0.44**	0.42**
	(0.016)	(0.032)
Error		-0.02***
		(0.000)
Rank Guess: 2		-0.51**
		(0.027)
Rank Guess: 3		-0.86***
		(0.004)
Rank Guess: 4		0.43
		(0.510)
Constant	0.11	0.45
	(0.384)	(0.370)
Controls	No	Yes
N	1,106	1,106

Notes: Dependent variable is *Guess Sent* (dummy variable equal to 1 if the sender sent their guess to the receiver in a given round and 0 otherwise). Control variables are *Male* (dummy variable equal to 1 for men and 0 for women), *Error* (normalized error of the sender in a given round), *Rank Guess* (indicator variables for subjects' self-confidence, takes values between 1-4). Column (2) also includes controls for period, risk preference, education, employment, and age, whose coefficients are available in Appendix Table B.4. Errors are clustered at the individual level. p-values are reported in parentheses; * p<0.1, ** p<0.05, *** p<0.01.

Since men and women senders in the Advice Treatment have similar performance levels for difficult questions (average normalized error is 11.62% and 11.59% for male and female senders, respectively, with $p = 0.942$), the underlying reason for the gender gap is not actual performance differences. Table B.3 presents demographics, risk attitudes, and self-confidence of senders by gender. In line with the literature, women are more risk-averse than men (Eckel and Grossman, 2008) and have lower self-confidence (Beyer, 1990; Niederle and Vesterlund, 2007). Column (2) of Table 2.2 shows that performance and self-confidence are two significant factors affecting the decision to send advice: subjects with lower performance (as measured by their normalized error) and lower self-confidence (as measured by their self-perceived rank of

guess accuracy) are less likely to send their guess.[14] However, gender still has a significant effect after controlling for sender's performance, self-confidence, and risk aversion as well as period, education, employment, and age. Hence, even though self-confidence and performance are both significant predictors of senders' decision to send advice, the gender gap in advice sending remains after controlling for these factors.

2.5.2 Does Gender Gap Persist When Senders Can Enforce Their Guess?

In the Dictator Treatment, the sender's guess is implemented as the pair's decision if the sender chooses to send it, independent of the receiver's answer. Contrary to the case when the senders' guess is simply advisory (i.e. the receiver chooses whether to implement it or not), there is no gender gap in guess-sending when senders can enforce their guess. There is no gender gap in senders' rate of guess-sending for any difficulty level in the Dictator Treatment ($p = 0.565$ for easy, $p = 0.265$ for medium, and $p = 0.478$ for difficult questions; see Figure B.2). Table 2.3 provides further evidence based on probit regressions relating guess sending to gender for a sample containing only difficult questions in the Dictator Treatment, clustering standard errors at the individual level. The results illustrate that performance (as measured by normalized error) and self-confidence (as measured by rank guess) are again significant predictors in a sender's decision to send their guess; however, contrary to the Advice Treatment, the gender coefficient is no longer significant in the Dictator Treatment with or without controls.

[14] It is not possible to make a meaningful inference about the coefficient on the lowest value of self-confidence in either treatment. As in Niederle and Vesterlund (2007), the number of subjects who guess that they are the worst in their group is very small (in both treatments, 2 out of 112 senders guessed their rank to be 4). The results are robust to excluding these subjects from the regression analysis as in Niederle and Vesterlund (2007).

Table 2.3: Probit Regressions Relating Advisor's Guess-Sending to Gender in Dictator Treatment for Difficult Questions

Advice Sent	(1)	(2)
Male	0.12	0.03
	(0.410)	(0.820)
Error		-0.02***
		(0.000)
Rank Guess: 2		-0.82***
		(0.000)
Rank Guess: 3		-1.06***
		(0.000)
Rank Guess: 4		-0.56
		(0.404)
Constant	-0.29***	0.69*
	(0.005)	(0.063)
Controls	No	Yes
N	1,106	1,106

Notes: Dependent variable is *Advice Sent* (dummy variable equal to 1 if the sender sent their guess to the receiver in a given round and 0 otherwise). Control variables are *Male* (dummy variable equal to 1 for men and 0 for women), *Error* (normalized error of the sender in a given round), *Rank Guess* (indicator variables for subjects' self-confidence, takes values between 1-4). Column (2) also includes controls for period, risk preference, education, employment, and age, whose coefficients are available in Appendix Table B.5). Errors are clustered at the individual level. p-values are reported in parentheses; * p<0.1, ** p<0.05, *** p<0.01.

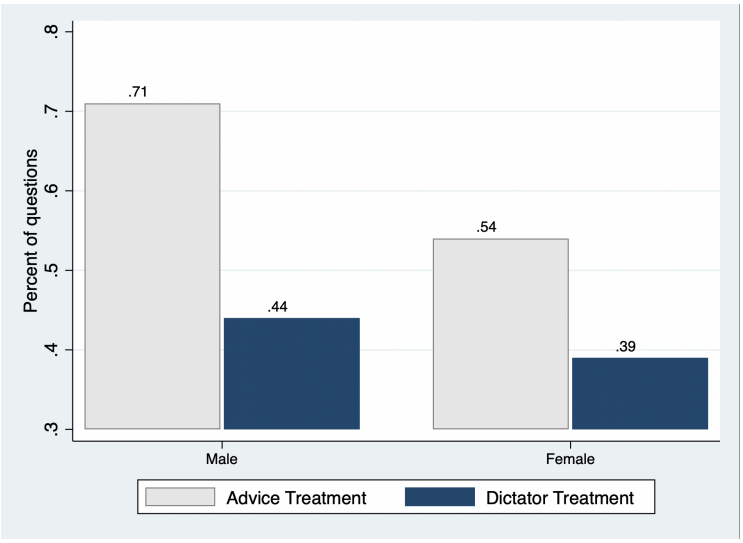
2.5.3 How Does Sender Behavior Change Across Treatments?

This section compares sender behavior across treatments by gender to better understand how the gender gap in guess sending shrinks when advice becomes enforceable. Figure 2.3 plots the percentages of difficult questions for which the senders send their guess, broken down by treatment and gender. Both male and female senders send their guess at significantly lower rates in the Dictator Treatment: the frequency of guess sending decreases from 71% to 44% for male senders ($p < 0.001$) and from 54% to 39% for female senders ($p = 0.024$). Note that the drop in frequency is larger for men both in terms of magnitude and significance.

Note that Figure 2.3 omits easy and medium questions. I find that differences in senders'

guess-sending frequency across treatments are observed only in difficult questions. Figure B.3 shows the percentages of easy, medium, and difficult questions for which the senders send their guess, broken down by treatment. Senders send their guess in 62% of the difficult questions in the Advice Treatment compared to 41% in the Dictator Treatment and this difference between treatments is statistically significant ($p < 0.001$). Figure B.4 plots the cumulative distribution functions (CDFs) of guess-sending frequency for difficult questions across treatments. The CDF in the Advice Treatment first order stochastically dominates the CDF in the Dictator Treatment.[15] Results show that both men and women send their guess less frequently (and the decrease is greater for men) in the Dictator Treatment for difficult questions.

Figure 2.3: Percentages of Difficult Questions For Which Senders Send Their Guess Across Treatments



Notes: Figure illustrates percent of difficult questions that the senders sent their guess, broken down by treatment and gender. The p-values for the difference of percentages between the Advice and Dictator Treatments are $p < 0.001$ for male senders and $p = 0.024$ for female senders.

[15] Based on Somers' D statistic (Newson et al., 2001), a randomly chosen sender in the Dictator Treatment is 38% less likely to send their guess for a difficult question than a sender in the Advice Treatment, with a 95% confidence interval of $[-0.540, -0.214]$.

2.5.4 Does Advice Improve Decisions?

Even though the main focus of this paper is on advice giving, I also examine the effect of advice on decision making for completeness and for relating the results to the existing literature. In order to explore the effect of advice on decision making, I examine receivers' performance in rounds with and without advice in the Advice Treatment. The receivers in the Advice Treatment received advice in 94%, 86%, and 60% of the rounds for easy, medium, and difficult questions, respectively. Since there is a high frequency of advice in easy and medium questions, this section focuses on the effect of advice for difficult questions.

In the Advice Treatment, the receivers' average normalized error for difficult questions in rounds with advice is 8.7%, while it is 13.7% in rounds without advice. Receivers perform significantly better in rounds for which they receive advice ($p < 0.001$). Note that if the senders' decision to send their guess is correlated with how hard the question is (assuming that some difficult questions are perceived harder than others), the performance difference in rounds with and without advice could be driven by omitted variable bias rather than advice improving decision making. To test this explanation, I examine how the receivers' performance in the Dictator Treatment is correlated with the senders' guess-sending decision. In the Dictator Treatment, rounds in which the sender sends their guess and the rounds in which they don't are indistinguishable from a receiver's perspective, as the receivers submit their guesses without any external information. If the senders' decision to send their guess were a proxy for question difficulty, one would expect the receivers in the Dictator Treatment to have lower performance in rounds for which the senders don't send their guess. Table B.6 reports the results of OLS regressions of receivers' performance on difficult questions on an indicator for whether the sender sends their guess for

each treatment, clustering standard errors at the individual level. In line with the earlier finding, receivers perform better in rounds for which the sender sends their guess (significant negative coefficient on normalized error, $p < 0.001$) in the Advice Treatment. However; the effect is not significant in the Dictator Treatment $(p = 0.410)$, which can be seen as suggestive evidence that endogeneity is not the driver of the performance gap.[16] In line with the advice literature, (e.g., Schotter, 2003; Cooper and Kagel, 2016) these findings support that presence of advice increases performance.

Next, I analyze whether there is a gender difference in how advice affects receivers. Contrary to senders in both treatments and receivers in the Dictator Treatment, male receivers perform significantly better than women in all difficulty levels in the Advice Treatment.[17] The average normalized error of men (women) is 9.6% (11.5%) for difficult questions, 2.5% (4.0%) for medium questions, and 0.5% (1.2%) for easy questions. The gender difference in performance is significant for all difficulty levels (p-values are 0.025, 0.020, and 0.028 for difficult, medium, and easy questions, respectively). Since there was no gender difference in the performance of senders in either treatment nor of receivers in the Dictator Treatment, the gender difference in the receivers' performance in the Advice Treatment means that presence of advice has a differential effect on men and women.

Why do male receivers outperform women in the presence of a sender in a task that is gender-neutral when subjects submit their answers without any external information? One possibility is that men incorporate advice better than women, in which case the gender gap should

[16] Since the decision to send advice differs between the Advice treatment and the Dictator treatment, endogeneity concern cannot be fully ruled out.

[17] Note that this finding does not contradict with the earlier result on gender-neutrality of the task. The receivers in the Advice Treatment submit their guesses after observing the advice (or observing that the sender did not send advice), which can affect their performance. I evaluate the gender-neutrality of the task based on the subjects' performance when they do not receive any external information.

arise only in rounds for which advice is sent. Figure B.5 illustrates that the difference in average normalized errors between men and women (8.3% and 9%, respectively) is not statistically significant ($p = 0.450$) for the rounds in which the receiver receives advice. On the contrary, the difference is significant in rounds without advice: men's average normalized error is 12%, compared to 16% for women ($p = 0.010$). The findings suggest that the gender gap in performance is driven by rounds without advice, falsifying the hypothesis that men incorporate advice better. One explanation could be that female receivers pay less attention to the task in the presence of a sender, relying on advice more than men. Alternatively, it could be that female receivers get intimidated by the difficulty of the question more than men when they do not receive advice, which disproportionately affects their performance in rounds without advice. It would be an interesting next step to investigate the underlying reasons of this differential effect of advice on men and women.

To investigate who gives better advice in difficult questions, I compare the average normalized error of male and female senders conditional on sending their guess. I do not find any gender difference in the quality of advice within treatments: In the Advice Treatment, the average normalized error of male and female senders are 10.4% and 9.6%, respectively ($p = 0.531$); while in the Dictator Treatment, it is 7.4% and 8.9% ($p = 0.346$).[18] One interesting finding is that conditional on sending their guess, men send more accurate guesses in the Dictator Treatment than in the Advice Treatment (average normalized error 10.4% in the Advice Treatment vs 7.4% in the Dictator Treatment, $p = 0.026$); while women do not increase the quality of their guess when the advice becomes enforceable (average normalized error 9.6% in the Advice Treatment

[18] In the remainder of this subsection, p-values are obtained using a t-test and clustering errors at the individual level.

vs 8.9% in the Dictator Treatment, $p = 0.709$). This is in line with the previous conjecture that men change their guess sending behavior when the advice becomes enforceable, while women behave similarly regardless of enforceability of advice.

Finally, I examine whether being matched to a male or a female sender leads to better answers by the receivers. I find no significant difference in average normalized error nor in payoff by the sender's gender in either treatment. In the Advice Treatment, the average normalized error of receivers with male and female senders in difficult questions are 10.3% and 10.7%, respectively ($p = 0.551$); while in the Dictator Treatment, it is 11.1% and 11.0% ($p = 0.947$). The results suggest that the gender difference in advice giving does not translate into receivers performing significantly better or worse depending on the sender's gender in the context of this experiment. The results are similar if I compare payoffs instead of average normalized errors. It is possible that given the percentage point gender difference in advice sending, the effect of receiving advice in difficult questions does not cause a big enough shift in performance to be significant in this task. I cannot rule out the possibility that in a different task in which the effect of advice on performance is larger, having a woman advisor who is less likely to send advice than a man might hurt earnings.

One caveat of the analysis regarding receivers' behavior is that receivers were matched with different senders in each round, so this may create a concern for non-independence across individual observations. Given the online nature of the experiment and the constraints associated with subject dropouts, it was not possible to do perfect random stranger matching. Since the focus of the paper was on senders' behavior, this was an intentional trade-off given the constraints. Even though the results of this subsection regarding receivers' behavior should be interpreted with this shortcoming in mind, receivers were randomly matched with a new sender in each round, so

any difference in their matches at the individual level is expected to cancel out when the results are compared at the gender or treatment level. The analysis regarding senders is not affected by this concern, since senders did not have any information about receivers' gender (or any other characteristics) at the time of sending their guess; hence there was no way for senders to send systematically different answers to men versus women receivers.

2.6 A Model of Guilt and Responsibility

Starting with Charness and Dufwenberg (2006) and Battigalli and Dufwenberg (2007), guilt aversion has been widely studied in a variety of settings. Settings in which guilt aversion has been shown to be relevant include deception (Battigalli et al., 2013), voting and public good games (Rothenhäusler et al., 2018), and cooperation (Peeters and Vorsatz, 2021). Guilt is a potential mechanism affecting subjects' behavior in this experiment; subjects may feel guilt from sending a misleading guess. Even though subjects do not learn the accuracy of their guess in the experiment, it is reasonable to assume that the time limit was insufficient to count the exact number of red balls in difficult questions, so that the senders do not anticipate making error-free guesses in these questions.

The intuition of a simple guilt model is that an agent feels guilt if they disappoint others. The amount of guilt an agent feels depends on their guilt sensitivity as well as the magnitude of disappointment they cause. In the model I introduce in this section, an agent's guilt sensitivity is related to how responsible they feel from others' payoffs. In most of the experimental guilt literature, one agent (whose guilt is in question) is clearly responsible for the outcomes of both players. For example, in the seminal Charness and Dufwenberg (2006) experiment, the agent

chooses whether to shirk or put in effort, which in turn determines both their and the principal's earnings. In case the outcome of subjects are determined as a result of both subjects' actions, it is possible for a subject to have lower guilt sensitivity through shifting the blame. Bartling and Fischbacher (2012) develop a measure of responsibility which assesses how responsible a subject is from another player's outcome when they delegate choosing between a fair and unfair allocation. Using this intuition, I develop a simple model of guilt and responsibility in which how responsible an agent feels from their pair's earnings affect their guilt sensitivity.

Advice Treatment

Consider a sender S whose guess for the number of red balls in the box is x_S and a receiver R whose guess is x_R. Let the earnings associated with answers x_S and x_R be m_S and m_R, respectively, if they were to be implemented as the final earning of the pair. Denote α as the probability that the sender's answer is implemented as the pair's final answer regardless of their choice to send advice (in the experiment, $\alpha = 0.05$).

If the sender sends their guess as advice, they potentially affect the answer of their group. Denote the sender's expectation of the answer that gets submitted as the pair's final answer as x_{SR} in the case that sender sends advice such that $x_{SR} = \beta_S x_S + (1 - \beta_S)x_R$, where $\beta_S \in [0, 1]$. β_S is the belief of the sender on how influential their answer is. In the extremes, if $\beta_S = 0$, the sender believes that their answer has no affect on the receiver's final answer, and if $\beta_S = 1$, the sender believes that the receiver follows the sender's advice verbatim. Given the payoff structure used in the experiment, the payoff associated with the answer will be $m_{SR} \in [min\{m_R, m_S\}, max\{m_R, m_S\}]$.

The guilt that a sender feels from sending their guess depends on how much they believe they disappoint the receiver as well as their guilt sensitivity. In the context of the experiment, I interpret the guilt sensitivity as the combination of how responsible the sender feels for the group's earnings if they send their guess to the receiver and a guilt intensity parameter. Let ρ denote the probability that the sender's answer is implemented as the group's final answer if they choose to send their guess[19]. Note that $1-\rho$ can be thought of as the scope for shifting the blame for the group's earnings to the receiver upon sending one's guess. For example, if $\rho = 0$ and a sender sends their guess, the pair's final answer is certainly determined by the receiver's guess. Hence, the sender can potentially shift all the blame to the receiver for their pair's final earnings. If $\rho = 1$, on the other hand, when a sender sends their guess, the pair's final answer is certainly determined by the sender's guess. So, there is no room for shifting the blame to the receiver upon sending one's answer. With this intuition, I denote the sender's guilt sensitivity, $G_S(a;\rho,d)$ as follows:

$$G_S(a;\rho,d) = \begin{cases} 0 & \textit{if a=not and d=R} \\ w_S \times \Gamma_S(\rho) & \textit{if a=send and d=R} \\ w_S & \textit{if d=S} \end{cases} \tag{2.1}$$

where $a \in \{\text{send,not}\}$ is the action of the sender, w_S is the guilt intensity parameter, $\rho \in [0,1]$ is the probability that the sender's guess is implemented as the pair's final answer if they send it, $d \in \{\text{S,R}\}$ denotes the role of the subject whose answer gets implemented as the pair's final answer, $\Gamma_S(\rho) \to [0,1] \times [0,1]$ and $\frac{d\Gamma_S(\rho)}{d\rho} \geq 0$.

[19] In the Advice Treatment, $\rho = \alpha$, while in the Dictator Treatment, $\rho = 1$

The functional form of $G_S(a;\rho,d)$ indicates that the sender does not feel any responsibility from their pair's earning when they don't send their guess and the receiver's answer gets implemented, that their sense of responsibility is a non-decreasing function of the probability that their answer is the one that counts when they send their guess but the receiver's answer gets implemented, and that they feel fully responsible for their pair's earnings when their answer is implemented as the final answer.

The utility of a sender S in the Advice Treatment upon sending their guess is represented by:

$$\begin{aligned} u^{A}_{S,send}(m_S,m_R,\alpha,\beta_S) =& (1-\alpha)\times[m_{SR}(\beta_S) - w_S \times \Gamma_S(\alpha)\times(m_{max}-m_{SR}(\beta_S))] \\ &+ \alpha\times[m_S - w_S\times(m_{max}-m_S)] \end{aligned} \tag{2.2}$$

where the first term corresponds to the case in which the sender's guess is sent as advice, in which case the sender expects the pair's monetary earnings to be $m_{SR}(\beta_S)$, has a guilt sensitivity of $w_S \times \Gamma_S(\alpha)$, and expects to disappoint the receiver by $m_{max} - m_{SR}(\beta_S)$; and the second term corresponds to the case in which the sender's guess is implemented as the final answer of the pair, in which case the sender expects the pair's monetary earnings to be m_S, feels full responsibility for the pair's earnings since their answer is implemented as the group's final answer so their guilt sensitivity is w_S, and expects to disappoint the receiver by $m_{max} - m_S$.[20]

[20] In a typical guilt model with two agents a la Battigalli and Dufwenberg (2007), agent 1 feels guilt from disappointing agent 2, which is calculated using the second order belief of agent 1 on what they think agent 2 expects to earn. Allowing for second order beliefs in this context leads to rationalisability of any action, in which case the model becomes too general to have any predictive power. For this reason, I make the simplifying assumption that the player 1 (sender) believes that the player 2 (receiver) is disappointed whenever the pair is unable to earn the maximum possible payoff in the experiment.

The utility of a sender S in the Advice Treatment if they don't send their guess is represented by:

$$u^{A}_{S,not}(m_S, m_R, \alpha, \beta_S) = (1-\alpha) \times m_R + \alpha \times [m_S - w_S \times (m_{max} - m_S)] \tag{2.3}$$

where the first term corresponds to the case in which the receiver's guess is implemented as the final answer of the pair, in which case the sender expects the pair's monetary earnings to be m_R, and does not feel responsible for the pair's earnings since the receiver's answer is implemented; and the second term corresponds to the case in which the sender's guess is implemented as the final answer of the pair, in which case the sender expects the pair's monetary earnings to be m_S, feels full responsibility for the pair's earnings since their answer is implemented as the group's final answer so their guilt sensitivity is w_S, and expects to disappoint the receiver by $m_{max} - m_S$.

Then, a sender would find it optimal to send their guess in the Advice Treatment if:

$$\frac{m_{SR}(\beta_S) - m_R}{m_{max} - m_{SR}(\beta_S)} \geq w_S \times \Gamma_S(\alpha) \tag{2.4}$$

Dictator Treatment

In the Dictator Treatment, when a sender sends their guess x_S, it is implemented as the final answer of the group. Hence, there is no room for beliefs on how influential the sender believes they are. The utility of a sender S in the Dictator Treatment upon sending their guess is

represented by:

$$u^{D}_{S,send}(m_S, m_R, \alpha, \beta_S) = m_S - w_S \times (m_{max} - m_S) \tag{2.5}$$

since the sender's guess is implemented as the pair's final answer for sure if they decide to send it.

The utility of a sender S in the Dictator Treatment if they don't send their guess is represented by:

$$u^{D}_{S,not}(m_S, m_R, \alpha, \beta_S) = (1-\alpha) \times m_R + \alpha \times [m_S - w_S \times (m_{max} - m_S)] \tag{2.6}$$

where the first term corresponds to the case in which the receiver's guess is implemented as the final answer of the pair, in which case the sender expects the pair's monetary earnings to be m_R, and does not feel responsible for the pair's earnings since the receiver's answer is implemented; and the second term corresponds to the case in which the sender's guess is implemented as the final answer of the pair, in which case the sender expects the pair's monetary earnings to be m_S, feels full responsibility for the pair's earnings since their answer is implemented as the group's final answer, and expects to disappoint the receiver by $m_{max} - m_S$.

Then, a sender would find it optimal to send their guess in the Dictator Treatment if:

$$\frac{m_S - m_R}{m_{max} - m_S} \geq w_S \tag{2.7}$$

Discussion

The main findings of this paper are that women are significantly less likely than men to send their guess as advice in difficult questions. Moreover, there is no significant gender difference in guess sending when the sender can enforce their guess as the pair's final answer, which is driven by men decreasing their guess sending in difficult questions compared to the case in which the guess was advisory. The simple model of guilt and responsibility introduced in this section aims to propose possible mechanisms that can lead to these findings.

i. Gender difference in guilt through feeling responsible for the pair's earnings: When there is room to shift the blame to another person, if women are more likely than men to feel responsible, we would have $\Gamma_S^{women}(\alpha) > \Gamma_S^{men}(\alpha)$. There are not any studies specifically focusing on gender differences in blame shifting, but Erat (2013) finds that women are more likely to delegate the responsibility for misleading another player than men, which would be in line with women feeling a higher sense of responsibility for the other player's earnings. If this is the case, it would make women less likely to send their guess compared to men in the Advice Treatment, since a higher sense of responsibility increases the threshold needed in Equation (2.4) to make guess sending more desirable than not sending. The gender difference in sense of responsibility would not lead to a gender difference in guess sending in the Dictator Treatment, since there is no room to shift the blame in this treatment and $\Gamma_S(.)$ does not affect the guess sending decision, as can be seen in Equation (2.7).

ii. Gender difference in guilt through the guilt intensity parameter: Gender difference in the guilt intensity parameter w_S, is not likely to be the mechanism driving the gender differences in advice giving. Note that the guilt intensity parameter affects both Equations (2.4)

and (2.7), and would have caused a gender difference in both treatments, not just in the Advice Treatment. In fact, there are several studies which find that men exhibit higher guilt aversion than women (Nihonsugi et al., 2022, Di Bartolomeo et al., 2022).[21] The guilt in these studies is through lying or breaking promises. It is not clear whether the gender difference in guilt aversion associated with breaking promises carries over to other settings such as the one in this experiment. Either way, gender differences in guilt intensity cannot explain the findings of this paper.

iii. Gender difference in beliefs about the influence on the receiver's answer: Conditional on believing that their answer will not strictly decrease their pair's earnings compared to the case if they don't send their guess, a higher belief in sender's influence on the receiver's answer would make them more likely to send their guess.[22] Manian and Sheth (2021) show that subjects don't expect women's advice to be followed as much as men. If women believe that their advice will not be followed as much as men, this would lead to $\beta_S^{women} < \beta_S^{men}$. For senders who expect same earnings from their answer ($m_S^{women} = m_S^{men}$), the gender gap in senders' belief on their influence would lead to $m_{SR}(\beta_S^{women}) < m_{SR}(\beta_S^{men})$, making the left-hand side of the Equation (2.4) lower for women. This would in turn make women less likely to send their guess in the Advice Treatment. Since β_S is not relevant for the pair's earnings in the Dictator Treatment, a gender difference in beliefs about how much one's advice will be followed would not lead to a gender difference in guess sending in this treatment.

iv. Gender difference in self confidence: Having lower confidence in accuracy of one's answer (m_S) would make it less desirable to send advice. Women are shown to have lower self-

[21] In both settings, the agent whose guilt is measured is fully responsible for both players' earnings, hence the relevant guilt sensitivity parameter in these studies is w_S.

[22] This conjecture relies on the assumption that the earning increase is relative to what the pair would have earned with the receiver's answer absent the advice, but the disappointment is relative to maximum possible earnings that the pair could achieve. If the assumption on senders believing that receivers expect to earn the highest possible amount fails, this mechanism becomes less plausible.

confidence than men (in this paper and also in others, e.g. Barber and Odean, 2001; Niederle and Vesterlund, 2007), so self-confidence is a candidate mechanism contributing to the gender gap in advice giving. If $m_S^{women} < m_S^{men}$, left-hand side of Equation (2.4) would be lower for women, making them less like to send their guess. However, there are two reasons against the gender differences in self-confidence being the main driver behind gender differences in advice giving. Firstly, even though I do not have confidence in answers at the question level, the gender difference in advice giving persists after controlling for self-confidence at the subject level. Secondly, all else equal, $m_S^{women} < m_S^{men}$ would also affect Equation (2.7), leading to lower guess sending by women in the Dictator Treatment, which I do not observe in the data. Hence, gender differences in self-confidence is unlikely to be the main driver of the gender differences in advice sending.

The experimental design of this study was intentionally kept simple to investigate whether there is a gender difference in advice giving in a controlled environment and if so, what its relation to difficulty of the task is. Based on findings in the previous literature suggesting that women may not like deciding on behalf of others more than men (e.g. Ertac and Gurdal, 2012 and Erat, 2013), I designed the Dictator Treatment to test whether this was the mechanism contributing to the gender gap in advice giving, and ruled it out as the underlying mechanism. However, the Dictator Treatment is not the proper control to identify which of the above guilt-related mechanisms lead to the gender differences in advice giving. The model introduced in this section aims to serve as a guide for further investigation to identify the underlying mechanism in future work.

2.7 Robustness Analysis

In this section, I conduct several robustness analysis to ensure that the main result of women being less likely to send advice than men in difficult questions is not driven artificially by the specific cutoffs chosen to classify questions' difficulty.

2.7.1 Varying the Cutoff for Classifying a Question as Difficult

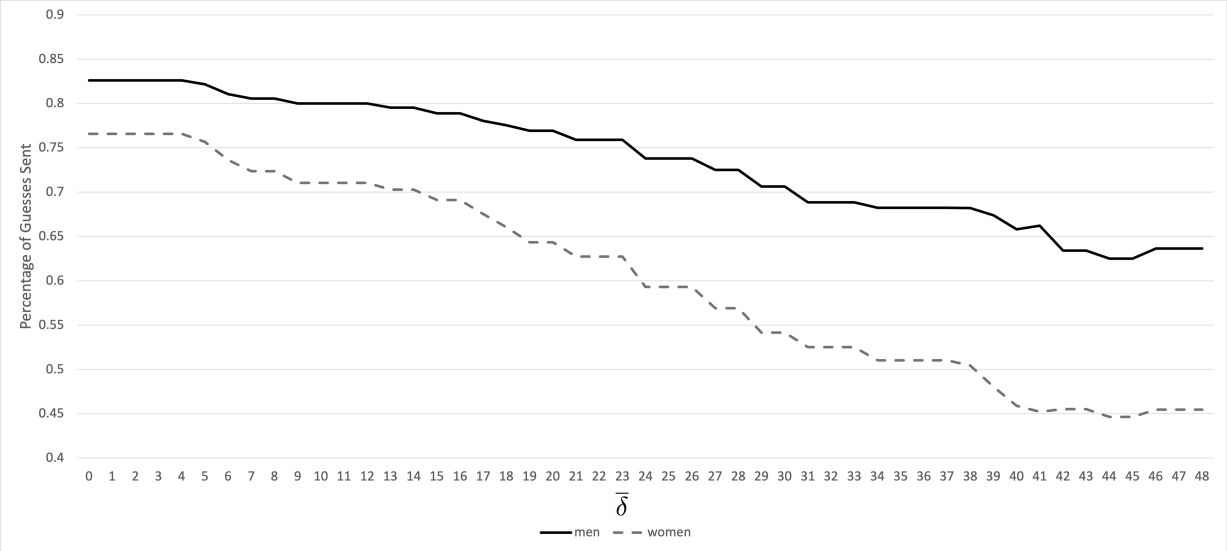
Consider a difficulty index, $\delta = (100 - |b - r|)/2$, where b and r correspond to the number of red and blue balls in the box, respectively. Note that a higher δ corresponds to a case in which the number of red and blue balls are closer to each other; hence, to a more difficult counting task. Denote $\overline{\delta}$ as the cutoff such that questions with $\delta > \overline{\delta}$ are classified as "difficult" questions. In the main analysis, I used $\overline{\delta} = 30$ as the cutoff to classify questions as "difficult". In this section, I vary the $\overline{\delta}$ cutoff from 0 (questions with $\delta > 0$, i.e. all questions, are classified as difficult) to 48 (questions with $\delta > 48$, i.e. only the question with 49 red balls and 51 blue balls is classified as difficult) and I report the percentage of advice sent by men and women in difficult questions based on this new definition of "difficult questions".[23]

Figure 2.4 illustrates the the percentage of questions for which senders send their guess in difficult questions in the Advice Treatment, broken down by gender, for different values of $\overline{\delta}$ used as a cutoff to classify whether a question is "difficult". The case where $\overline{\delta} = 0$ is equivalent to investigating the gender difference in advice sending without breaking the analysis down by difficulty, since all questions are classified as "difficult" in this case. The figure depicts that for

[23]The reason for varying $\overline{\delta}$ up to 48 is that for higher values of $\overline{\delta}$, no question can be classified as difficult, since the question with the closest number of red and blue balls in the experiment was 49 red balls and 51 blue balls.

all cutoff values $\overline{\delta}$, men send more advice than women in difficult questions, and the gender gap in advice sending increases as the threshold for question difficulty increases. Table B.7 shows how many questions are classified as "difficult" for each possible value of $\overline{\delta}$, along with the percentage of advice sent by men and women, the gender difference in percentage of advice sent, and the p-value associated with the gender difference in advice sending when question difficulty is determined by the corresponding $\overline{\delta}$. The gender gap in advice sending becomes significant at the 5% level when all but the easiest 4 questions are classified as "difficult" (when $\overline{\delta} = 7$, 21 out of 25 questions are classified as "difficult"). The difference remains positive and mostly increasing for higher values of $\overline{\delta}$. The gender difference in advice giving for difficult questions remains significant for all cutoff levels except for when only one question remains to be classified as difficult (when $\overline{\delta} \geq 46$, only 1 out of 25 question is classified as "difficult"). The analysis in this subsection shows that women being less likely to send their guess as advice in difficult questions is robust to alternative cutoffs that can be used to determine question difficulty.

Figure 2.4: Percentage of Advice Sending in Difficult Questions by Gender for Different Values of $\overline{\delta}$



Notes: Figure illustrates the percentage of questions for which senders send their guess in difficult questions in the Advice Treatment, broken down by gender. The x-axis varies the $\overline{\delta}$ cutoff used to classify a question as "difficult". At $\overline{\delta} = 0$, all questions are classified as difficult. At $\overline{\delta} = 48$, only one question (with 49 red balls and 51 blue balls) is classified as difficult.

2.7.2 Regressions Controlling for the Difficulty Index as an Alternative to Breaking Down the Data by Categorical Difficulty Levels

I investigate the relationship between advice giving behavior and gender by controlling for difficulty of the question measured by δ (as defined in Section 2.7.1), as an alternative to analysing advice giving separately for each categorical difficulty level, which was the analysis conducted in Section 2.5.1.

Table B.8 reports the results of Probit regressions relating guess sending in the Advice Treatment to gender, difficulty index δ, and their interaction for all questions without analyzing each difficulty level separately. The coefficient of the interaction term, *Male*$\times$*Delta* is positive

and significant ($p = 0.001$), showing that men become significantly more likely than women to give advice as the question difficulty increases. The coefficient of the gender dummy being negative and marginally significant ($p = 0.055$) indicates that men send less advice than women in the easiest question. The coefficient of the difficulty index, δ, is negative and significant ($p < 0.001$), confirming that advice sending decreases as question difficulty, indicated by δ, increases. The results are similar after controlling for sender's performance, self-confidence, and risk aversion as well as period, education, employment, and age.

2.7.3 Regressions Interacting Gender with Difficulty Levels

Finally, I provide regression results interacting the *Male* indicator with *Medium* and *Difficult* indicators as a second alternative to analysing advice giving separately for each categorical difficulty level. Table B.9 reports the results of the Probit regressions relating guess sending in the Advice Treatment to gender and its interaction with each difficulty level.

The second half of Table B.9 reports the gender difference in advice sending for each difficulty level. For investigating the gender difference in difficult questions, I examine the difference of coefficients corresponding to advice sending in difficult questions by men ($\beta_{Constant} + \beta_{Male} + \beta_{Difficult} + \beta_{Male \times Difficult}$) and women ($\beta_{Constant} + \beta_{Difficult}$). The p-values are obtained using Chow-tests on equality of these coefficients to each other. I do a similar analysis for medium and easy questions. I find that men are significantly more likely than women to send advice in difficult questions ($p = 0.016$), that there is no gender difference in advice sending in medium questions ($p = 0.649$), and that women are more likely than men to send advice in easy questions ($p = 0.021$). The results are similar after controlling for sender's performance, self-confidence,

and risk aversion as well as period, education, employment, and age.

Hence, the main result that women are less likely than men to send advice in difficult questions is robust to pooling all questions together and controlling for difficulty level as an alternative to examining behavior in difficult questions in isolation. The advice sending behavior in medium questions is also similar. The only difference in findings is that women are significantly more likely than men to send advice in easy questions when questions are pooled, which is in line with the previous analysis in terms of the direction of the result; but, the finding was not significant when easy questions were examined in isolation.

2.8 Conclusion

This paper contributes both to the advice literature and to the literature that explores why women are underrepresented in high-profile positions in the labor market. Using a gender-neutral task for which the incentives of the sender and the receiver are perfectly aligned, I show that female senders are significantly less likely than men to send advice to the receiver for difficult questions. The gender gap in advice giving persists even after controlling for senders' performance, self-confidence, demographics, and risk preferences. On the other hand, when the senders choose whether to be their pair's decision maker rather than whether to send advice to the receiver of their pair, there is no significant gender gap in guess-sending. Both men and women send their guess significantly less in the Dictator Treatment, but the decrease is greater for men, diminishing the gender gap. I consider gender differences in guilt sensitivity through feeling responsible for the pair's earnings, gender differences in beliefs about affecting others' answer, and gender differences in self-confidence as possible mechanisms that can lead to gender differences in advice

giving documented in this paper.

This paper also supports the findings in the advice literature that the presence of advice increases performance of receivers (e.g., Schotter, 2003; Cooper and Kagel, 2016). I additionally find that the difficulty of the problem at hand is crucial for advice to have a significant effect on performance. Furthermore, even though I use a task in which there are no gender differences in performance for subjects who do not receive any external information (i.e. senders in both treatments and receivers in the Dictator Treatment), male receivers perform significantly better than women in the Advice Treatment. The performance gap is driven by rounds in which the receivers did not receive advice. Gender differences in attention in the presence of a sender and gender differences in discouragement upon observing no advice are some possible explanations for the performance gap in the Advice Treatment.

In this paper, I focus on subjects' willingness to give advice in isolation, to rule out potential confounds that could affect advice giving. A fruitful future direction for research is to investigate the effect of adding channels such as feedback about subject performance, whether advice was followed, and information about other subjects' gender to better understand the underlying mechanisms behind the documented gender gap. Moreover, familiarity of the subjects, reputation, potential for backlash, and the nature of the task may be relevant determinants affecting advice giving in certain setups. While these are interesting avenues meriting further study, they are not the focus of this project. The effect of these channels and welfare effects of advice giving are important questions left for future work.

Chapter 3: Evidence Games: Lying Aversion and Commitment

3.1 Introduction

In voluntary disclosure literature where there is an informed sender and an uninformed receiver, the case where the receiver moves first and commits to a reward policy corresponds to a mechanism setup and the optimal mechanism has been studied (see e.g. Bull and Watson, 2007; Deneckere and Severinov, 2008; Green and Laffont, 1986); alternatively, the case where the receiver decides on the reward after observing the sender's decision corresponds to a game setup and the equilibrium of the game has been studied (see e.g. Dye, 1985; Grossman, 1981; Grossman and Hart, 1980; Milgrom, 1981). The link between these two settings is an important question. Glazer and Rubinstein (2006) has studied a setting where commitment does not have any value, in other words, the outcome of the optimal mechanism could be obtained in the equilibrium of the game setup. This result has been extended and investigated in other settings (see e.g. Ben-Porath et al., 2019; Sher, 2011).[1] Particularly, Hart et al. (2017) extended this result to *evidence games*. A distinguishing feature of evidence games is that the sender's utility function is increasing in the reward independent of his type, the receiver's utility function depends on the sender's type and satisfies the single-peakedness condition. Furthermore, senders cannot lie about the pieces

[1] In the closely related Bayesian persuasion literature initiated by Kamenica and Gentzkow (2011), the informed sender has commitment power that can be used to persuade the uninformed receiver. See Fréchette et al. (2019) for an experimental analysis of subject behavior when senders have commitment power.

of evidence that they have, but they can choose not to disclose some pieces of their evidence. In this paper, we experimentally investigate whether commitment has any value in evidence games by testing the equivalence of the optimal mechanism and the game equilibrium outcomes.

Consider an agent (informed sender) who is asked to submit a self-evaluation for an ongoing project, and a principal (uninformed receiver) who decides on the agent's reward. If the agent conducted his part as planned, he does not have any evidence to report at this point. But if he made a mistake which can't be traced back to him unless he discloses it, he may choose to show his mistake or act as if he has no evidence.[2] The agent wants to have a reward as high as possible independent of his evidence but the principal wants to set the reward as close as possible to the agent's value. Our experimental setup mimics this motivating example and asks: Does it matter whether the principal commits to a reward policy and then the agent decides whether to reveal the evidence or not, or the principal moves second after observing the agent's decision?

First, to see the intuition of the theoretical result of why commitment does not have any value in evidence games, assume that with probability of 50%, the agent conducted his work without a mistake (High type) and his value is 100 but with probability of 50%, he made a mistake (Low type) and his value is 0. In the mechanism setup (where there is commitment), the only way the principal can separate low and high types is to set a higher reward for low evidence (which can only be disclosed by low types) than for no-evidence, which is suboptimal for the principal. The optimal mechanism with commitment is that the principal sets a reward of 50 for no evidence and a reward lower than or equal to 50 for low evidence; which implies that the optimal mechanism cannot separate low and high type agents. So, the unique outcome of the

[2] Alternatively, say a professor has submitted to a journal, and the dean, who decides on professor's salary increase, asks whether he got a desk rejection.

optimal mechanism is that both agents still get 50 payoff. In the game setup (where there is no commitment), there is a unique sequential equilibrium where the low type hides his evidence and pretends as if he is a high type. Since neither type discloses any evidence, the principal sets the reward at $50 (= 50\% \times 100 + 50\% \times 0)$. Hence, commitment does not have any value.

In the simple example above, which is based on our experimental setup, there is a unique equilibrium. However, in a general evidence-game setup, the types and the evidences can be quite rich, and there may be multiple equilibria. Hart et al. (2017) identify a refinement (*truth-leaning refinement*) such that in evidence games, the outcomes coincide for the truth-leaning equilibria without commitment and the optimal mechanism with commitment. Since our aim is to test whether commitment has any value in evidence games, the environment in the experiment needs to be simple enough so that it is not affected from subjects' ability to do Bayesian updating or equilibrium selection. Indeed, the equilibrium with and without commitment in the aforementioned example do not require the subjects' ability to do complex Bayesian updating. Furthermore, the equilibrium outcome is unique; hence, the equilibrium selection is not a concern. Therefore, this simple environment is ideal to test the value of commitment in evidence games, and we used it in our experiment.

Despite this simple setup, our experimental results yield that commitment actually makes a difference. Particularly, the principals who choose the reward after observing the agent's action behave in line with equilibrium predictions, while the principals who commit to a reward scheme in advance choose a reward strictly higher than the optimal reward for no evidence. We then theoretically show that such a divergence between outcomes in the presence and absence of commitment is explained by accounting for lying aversion. Finally, in line with a lying aversion model, we show that percent of agents who withhold their evidence varies across these two setups

and that when the agents move second, their decision to withhold evidence is affected by reward amounts even when there is no payoff gain from being truthful.

Lying aversion in games with strategic interactions has been well-documented in the literature.[3] Gneezy (2005) is the first one to experimentally measure people's aversion to lying in a sender-receiver game. His findings suggest that people sometimes act truthfully even if they have to forgo monetary payoffs to do so. Moreover, he shows that one's own earnings and the harm that lying causes to others are both important factors when deciding to lie. Sánchez-Pagés and Vorsatz (2007), Serra-Garcia et al. (2013), and Ederer and Fehr (2017) are other experimental studies documenting behavior consistent with truth-telling preferences in strategic environments. Along with the experimental papers presenting evidence for lying averse agents, there are also various theoretical papers incorporating aversion to lying in their models. Lacker and Weinberg (1989), Goldman and Slezak (2006), Guttman et al. (2006), Deneckere and Severinov (2017) are some examples studying optimal mechanism design with costly state misrepresentation. Kartik et al. (2007) and Kartik (2009) incorporate costly state misrepresentation to cheap talk setup of Crawford and Sobel (1982) and examine strategic communication with lying costs. Even though there is no study which explicitly accounts for lying averse agents in evidence games, according to the truth-leaning refinement of Hart et al. (2017), a sender prefers disclosing truthfully when the payoffs between disclosing the whole truth and withholding some evidence are equal. This refinement is justified by an infinitesimal increase in agent's utility for telling the whole-truth or equivalently by an infinitesimal decrease in agent's utility for withholding an evidence. We show that if this utility decrease for not revealing the whole truth is different than zero, even if it is

[3]Lying aversion has been widely investigated in setups that do not involve strategic interactions (see e.g. Abeler et al., 2019; Fischbacher and Föllmi-Heusi, 2013; Gneezy et al., 2018).

small, the outcome equivalence result in evidence games no longer holds.

With this in mind, consider an agent who bears a small but strictly positive cost of lying such that his utility is reduced by this cost if he doesn't reveal the whole truth. When the principal commits to a reward scheme in advance, say in the aforementioned example, the principal sets the reward equal to \$50 if the agent presents no evidence and equal to \$49 if the agent presents evidence for his type. If a low type agent's cost of lying is higher than the additional \$1 he would earn by lying, then the agent would actually disclose his evidence and get the lower payoff. We show that in the presence of agents with a strictly positive cost of lying, the optimal reward set for no evidence when the principal commits to a reward scheme is higher than the reward when the agents do not have a cost of lying. On the other hand, when there is no commitment, the equilibrium outcome remains unchanged even in the presence of lying averse agents, just as we observe in the data.

The remainder of this paper is organized as follows. Section 3.2 presents the model. Section 3.3 describes the experimental design and protocol and states the hypotheses under the standard model. Section 3.4 presents the experimental results. Section 3.5 introduces a model with lying averse agents and discusses how predictions of the model compare to the experimental findings. Section 3.6 concludes.

3.2 Model

Following the model of Hart et al. (2017), there is an agent denoted by A and a principal denoted by P. The agent has a type $t \in T$ where T is a finite set. The agent's type is chosen according to a probability distribution $q = (q_t)_{t \in T} \in \Delta(T)$ where $q_t > 0$ for all $t \in T$. The agent

knows his type, while the principal only knows the probability distribution of types. The agent has a value $v(t)$ associated with his type t.

Each type has access to a set of pieces of evidence $E_t \subseteq E$ where E is the set of all pieces of evidence. A type t agent can choose to reveal the whole truth (i.e. send E_t) or can send the evidence associated with a type who has less evidence (i.e. send $E_m \subseteq E_t$). In other words, agent can choose to withhold information. Denote the set of available messages to type t agent as $L(t) := \{m \in T : E_m \subseteq E_t\}$. In the general setting, the agent chooses a message $m \in L(t) \subseteq T$ to send to the principal, while the principal chooses a reward $x \in \mathbb{R}$ to send to the agent.

The agent's utility for reward x, $U^A(m, x; t) = u(x)$, does not depend on either the type t nor the message m. It is assumed to be a continuously differentiable and strictly increasing function. The principal's utility, $U^P(m, x; t) = w(x, v(t))$, depends on the reward and the value of the agent with type t but not on message m. We assume that the principal's utility is a continuously differentiable, strictly concave and single peaked function maximized at $x = v(t)$. These utility functions capture the idea that the agent wants as much reward as possible, while the principal wants the reward to match the value of the agent.

3.3 Experimental Procedures and Hypotheses

We conducted the experiment at the Experimental Economics Laboratory at the University of Maryland (EEL-UMD). We recruited 128 subjects from the University of Maryland's undergraduate student pool via ORSEE (Greiner, 2015). None of the subjects participated in more than one session. We used the experimental software zTree (Fischbacher, 2007) to design the experiment. We conducted 8 sessions with 16 subjects in each. There were equal number of subjects in

each treatment. Average session lasted about an hour and average payment was $\$15.4$, including the $\$7$ show-up fee. Payoffs in the experiment were in Experimental Currency Units (ECUs) with a conversion rate of 10 ECUs for \$1. Each session of our experiment consists of two parts. Paper instructions were distributed and read aloud prior to the start of each part. Before the experiment began, each subject was required to answer two questions that checked their understanding. If a subject failed to answer either of these questions correctly, they received a pop-up message informing them that they need to correct their relevant answer. The experiment started only after every subject answered these questions correctly. The instructions, sample screenshots and the two understanding questions are in Appendix C.4.

The first part of the experiment consisted of 20 independent periods. Each subject was assigned the role "agent" or "principal" in the first period, which remained fixed throughout the experiment.[4] In each period, subjects were randomly matched to another subject who was of the other role and played a single-shot game where the agent sends a message regarding their type and the principal chooses a reward between 0 and 100 for the agent.

The agent could be one of the two types: *high* or *low* with values 100 and 0, respectively. Each type occurred with probability 50%. Low type agents had evidence for their type, whereas high type agents did not have evidence. In order to ensure that the subjects understood the difference between "evidence" and "type", we used sentences associated with each type of agent to be sent as messages instead of letting the agents send $m \in \{\text{high, low}\}$. Low type agents had access to the messages $m \in \{$"My type is low", "I don't have evidence for my type"$\}$ whereas high type agents only had access to the message $m \in \{$"I don't have evidence for my type"$\}$.[5] The

[4] In the experiment, we stated the role of the agent as "sender" and the role of the principal as "receiver". We continue referring to the roles as "agent" and "principal" for the remainder of this paper for ease of reading.

[5] Information about agent types is summarized in Table C.1.

payoff of the agent was equal to the reward chosen by the principal. The payoff of the principal was $100 - |x - v(t)|$, where x is the reward that the agent receives and $v(t)$ is the true value of the agent of type t. Note that the principal's payoff is maximized when the reward is equal to the true value of the agent. The probability distribution of the agent's type, available messages for each type, and payoff functions for both roles were common knowledge to both agents and principals. All subjects knew that these information were common knowledge to both roles.

There were two treatments which differed in whether the agent [No-Commitment Treatment] or the principal [Commitment Treatment] was the first-mover. In the No-Commitment treatment, the agent chose which message to send to the principal among the messages that were available to his type. Once the agent chose which message to send, the principal observed the message and then chose a reward for the agent.[6] In the Commitment treatment, the principal chose a reward for each possible message that she could receive before she observed the message. The agent chose which message to send after observing the reward policy set by the principal. The type of the agent was randomly determined in each period.

In the second part of the experiment, identical in both of the treatments, we elicited the subjects' risk preferences and ability to do Bayesian updating using two incentivized activities. In the first activity, we asked the subjects to make choices in a menu of ordered lotteries following Holt and Laury (2002) to elicit their risk preference. In the second activity, following Charness and Levin (2005), we asked the subjects a Bayesian updating question which paid 10 ECUs if their answer was correct.

[6]For studies in which the off-equilibrium behavior is important, one may use a strategy method for the principal's decision. Since our aim is to investigate the outcome equivalence between treatments, it is sufficient to observe the on-equilibrium behavior, and hence we use the direct-response method as in many sequential game experiments (see Brandts and Charness, 2011 for a detailed survey on the strategy method).

Hypotheses

Based on the model in Section 2, given the material payoffs, the principal's utility is $w(100 - |v(t) - x|)$ and the agent's utility is $u(x)$. Let x_0 and x_- denote the reward set for no evidence and low evidence, respectively.

First, let's consider the No-Commitment (NC) setup. In the unique sequential equilibrium of the game, both low and high type agents send no evidence. If the principal were ever to observe low evidence, the best response would be to set the reward equal to 0 since the problem of the principal in this case is to choose $x_- \in [0, 100]$ to maximize $w(100 - x_-)$. So, $x_-^{NC} = 0$ in the No-Commitment setup. If the principal observes no evidence, the principal does not gain any new information out of this message since all low type agents will pretend to be high type as long as ($x_0 > 0$). The principal's problem upon observing no evidence is then to choose $x_0 \in [0, 100]$ to maximize $0.5 \cdot w(100 - x_0) + 0.5 \cdot w(x_0)$, which results in $x_0^{NC} = 50$ (Hypothesis 1).

Hypothesis 1 *The reward set for no evidence in the No-Commitment treatment is equal to 50.*

In the Commitment (C) setup, commitment does not help the principal. The only way to separate low type agents from high types is to set $x_- > x_0$ (since incentive compatibility constraint is $u(x_-) \geq u(x_0)$), which is not optimal since expected utility of the principal is decreasing in x_-. So, the problem of the principal is still to choose $x_0 \in [0, 100]$ to maximize $0.5 \cdot w(100 - x_0) + 0.5 \cdot w(x_0)$, which results in $x_0^C = 50$ and $x_-^C \leq 50$ (Hypothesis 2) in the optimal mechanism. Hence, commitment should not matter (Hypothesis 3).

Hypothesis 2 *The reward set for no evidence is equal to 50 in the Commitment treatment.*

Hypothesis 3 *The reward set for no evidence in the No-Commitment treatment is equal to the reward set for no evidence in the Commitment treatment.*

Next, we turn to the agents. Since high type agents do not have any evidence to disclose or withhold, we will look at the behavior of low type agents. In the unique sequential equilibrium of the No-Commitment treatment, if the agent reveals his evidence, the principal learns his type and gives $x_{-}^{NC} = 0$. However, if the agent withholds his evidence, the principal cannot learn his type and gives $x_{0}^{NC} = 50$. So, in the No-Commitment treatment, the low type agent always withholds his evidence to get a higher reward. Similarly, in the optimal mechanism of the Commitment treatment, the principal offers a higher reward for no-evidence, $x_{0}^{C} = 50$, than for low-evidence, $x_{-}^{C} \leq 50$, and the low type agent chooses not to reveal his type in any sequential equilibrium of the Commitment treatment. Hence, withholding evidence behavior should be identical in both treatments (Hypothesis 4).

Hypothesis 4 *The percentage of low type agents who withhold their low evidence in the No-Commitment and Commitment treatments are equal.*

Additionally, in the Commitment treatment, the agent is the second mover, so a low type agent decides whether to reveal his evidence or not after seeing the rewards committed by the principal. Unless the reward for low evidence is higher than the reward for no evidence, the reward amounts should not affect an agent's decision to withhold evidence. For example, say the reward for no evidence is 50. Then, whether the reward for low evidence is 49 or 0 should not affect the agent's decision. His decision solely depends on the highest reward rather than the amounts (Hypothesis 5).

Hypothesis 5 *In the Commitment treatment, provided that the reward for low evidence is not higher than the reward for no evidence, increasing or decreasing the reward amounts does not change the percentage of low type agents who withhold their evidence.*

3.4 Results

In this section, we report the experimental results on the reward for no evidence and low evidence set by the principals, truthful behavior of agents, and payoffs of subjects. We compare the results with the hypotheses discussed in the previous section.

Our data is independent at the session level, but there are 8 independent session clusters. Therefore, for more reliable inferences, throughout the analysis, we use non-parametric tests and wild cluster bootstrap method for the regression analysis (see Cameron et al., 2008). In particular, we follow the wild cluster bootstrap procedure of Cameron and Miller (2015) for OLS regressions and the score wild bootstrap procedure of Kline and Santos (2012) for tobit and probit regressions. We compute 95% confidence intervals and p-values by using the wild bootstraps algorithms developed by Roodman et al. (2019) with 9,999 bootstrap replications and clustering at the session level.[7]

We begin our analysis with the reward decision of the subjects in the role of a principal. First, we compute the average reward set for no-evidence and the average reward set for low-evidence in each treatment by all principals (see Table 3.1).

[7] All results are robust to conducting the regression analyses without the wild cluster bootstrap method. Tables obtained without the wild cluster bootstrap method are reported in Appendix C.2.

Table 3.1: Average Rewards by Treatment

Treatment	Reward for No Evidence	Reward for Low Evidence
No-Commitment	50.58	19.36
	[85.4%]	[14.6%]
	(593)	(47)
Commitment	60.42	27.05
	[72.2%]	[27.8%]
	(640)	(640)

Note: Percent of low type subjects who chose the corresponding message in each cell are reported in brackets, number of observations are reported in parentheses.

Reward for No Evidence:

Theoretically, the reward for no evidence should be equal to 50 in both No-Commitment and Commitment treatments. The experimental data shows that the average reward set by principals for no evidence is 50.58 in the No-Commitment treatment and 60.42 in the Commitment treatment (see Table 3.1). By using a Wilcoxon signed-rank test, we compare the estimated constant to the theoretical prediction.[8] We find that the reward for no evidence in the No-Commitment treatment is not significantly different than 50 ($p = 0.133$), which is in line with Hypothesis 1; yet it is significantly more than 50 in the Commitment treatment ($p < 0.001$), which falsifies Hypothesis 2. These results are robust when we condition to the reward set by subjects who are classified as risk averse ($p = 0.162$ in the No-Commitment treatment and $p < 0.001$ in the Commitment treatment).

[8]Unless otherwise stated, all p-values to compare distributions are obtained using the Mann Whitney U-test and all p-values to compare measures to benchmarks are obtained using the Wilcoxon signed-rank test in non-parametric analysis.

Result 1 *(a) In the No-Commitment treatment, the reward set for no evidence is not significantly different from the equilibrium reward. (b) In the Commitment treatment, the reward set for no evidence is significantly higher than the optimal reward.*

To measure treatment effects, we use a tobit regression relating reward for no evidence on the treatment dummy (depicted in Table 3.2). The coefficient of the commitment variable is positive and significant ($p = 0.005$), falsifying Hypothesis 3. Treatment variable remains significant after controlling for period, gender, risk attitudes, and ability to Bayesian update ($p = 0.002$).

Result 2 *The reward set for no evidence in the Commitment treatment is significantly higher than that in the No-Commitment treatment.*

Table 3.2: Tobit Regressions Relating Reward for No-Evidence to Treatment

	(1)	(2)
Commitment	15.32***	15.06***
	(0.005)	(0.002)
Period		-0.47
		(0.327)
Gender		-1.6
		(0.899)
Risk aversion		-0.89
		(0.885)
Ability to		-7.0
Bayesian update		(0.602)
Constant	50.3***	59.9***
	(0.003)	(0.003)
Observations	1,233	1,233

Notes: Dependent variable is *reward for no evidence*, bounded between 0 and 100. *Commitment* is a dummy variable that takes the value 1 if subject is in Commitment treatment and 0 if subject is in No-Commitment treatment. *Period* takes values from 1 to 20 and represents the period. *Gender* is a dummy variable that takes the value 1 if subject is female and 0 otherwise. *Risk Aversion* takes the value 1 if the subject is classified as risk averse based on the number of safe options they chose in Activity 1 and 0 otherwise. *Ability to Bayesian update* is a dummy variable that takes the value 1 if subject answered the Activity 2 question of Part II correctly and 0 otherwise. p-values computed by score wild bootstrap procedure are in parentheses (clusters are at the session level); * p<0.1, ** p<0.05, *** p<0.01.

Reward for Low Evidence:

In the Commitment treatment, as expected, the reward for low evidence is rarely higher than the reward for no evidence (only 3.9%). For each policy, we take the difference between the reward for no evidence and the reward for low evidence. We find that that this difference is significantly higher than 0 ($p < 0.001$). Additionally, the average reward set by principals for low evidence is 27.05 and it is significantly less than 50 ($p < 0.001$) but significantly more than 0 ($p < 0.001$). On the other hand, in the No-Commitment treatment, observing low evidence is an off-equilibrium behavior. As expected, when the principals observe low evidence, 59.57% of the reward for low evidence is equal to 0 in the No-Commitment treatment.

Withholding Information:

Next, we examine the percentage of subjects withholding their information (i.e. sending no evidence when they are low type) across treatments. In the No-Commitment treatment, 85.4% of low type subjects withhold their low evidence, while this ratio is 72.2% in the Commitment Treatment. These percentages are significantly different than one another ($p < 0.001$, both with a test of proportions and with a Mann–Whitney test), falsifying Hypothesis 4. The difference in withholding information across treatments may be due to the principal's reward choice or due to the agent's behavior. In the Commitment treatment, if a principal sets the reward for low evidence strictly higher than the reward for no evidence, it becomes optimal even for a payoff maximizing low type agent to reveal his evidence. Even when we exclude those rare cases, the percentage of low type agents withholding their evidence in the Commitment treatment (74.4%) is still significantly lower ($p < 0.001$).

Result 3 *The subjects with low evidence are significantly less likely to withhold their evidence in the Commitment treatment than those in the No-Commitment treatment.*

To test Hypothesis 5, we use a probit regression relating withholding information of low type agents to the rewards for no evidence and low evidence in the Commitment treatment conditioning on the cases in which the reward for low evidence is not higher than the reward for no evidence. Table 3.3 shows that agents are more likely to withhold evidence when reward for no evidence is higher ($p = 0.012$); yet, they are less likely to withhold evidence when the reward for low evidence is higher ($p = 0.013$), falsifying Hypothesis 5. Reward for no evidence and reward for low evidence both continue to have a significant effect on propensity to withhold evidence after controlling for period, gender, risk attitudes, and ability to Bayesian update ($p = 0.011$ and $p = 0.017$, respectively).[9] To test Hypothesis 5, we use a probit regression relating withholding information of low type agents to the rewards for no evidence and low evidence in the Commitment treatment conditioning on the cases in which the reward for low evidence is not higher than the reward for no evidence. Table 3.3 shows that agents are more likely to withhold evidence when reward for no evidence is higher ($p = 0.012$); yet, they are less likely to withhold evidence when the reward for low evidence is higher ($p = 0.013$), falsifying Hypothesis 5. Reward for no evidence and reward for low evidence both continue to have a significant effect on propensity to withhold evidence after controlling for period, gender, risk attitudes, and ability to Bayesian update ($p = 0.011$ and $p = 0.017$, respectively).[10]

[9] Additionally, we report the results of a probit regression relating withholding information of low type agents to the difference between rewards in Table C.2. The difference between the reward for no evidence and reward for low evidence has a significant effect on low type agents' propensity to withhold evidence in the Commitment Treatment.

[10] Additionally, we report the results of a probit regression relating withholding information of low type agents to the difference between rewards in Table C.2. The difference between the reward for no evidence and reward for low evidence has a significant effect on low type agents' propensity to withhold evidence in the Commitment Treatment.

Result 4 *In the Commitment treatment, subjects with low evidence are significantly more likely to withhold evidence as the reward for no evidence increases and are significantly less likely to withhold evidence as the reward for low evidence increases, even when the reward for low evidence is not higher than the reward for no evidence.*

Table 3.3: Probit Regressions Relating Withholding Information to the Rewards in the Commitment Treatment Conditioning on the Difference being Positive

	(1)	(2)
Reward for	0.018**	0.019**
No Evidence	(0.012)	(0.011)
Reward for	-0.026**	-0.027**
Low Evidence	(0.013)	(0.017)
Period		0.019*
		(0.081)
Gender		-0.289
		(0.194)
Risk aversion		-0.871
		(0.173)
Ability to		0.067
Bayesian update		(0.865)
Constant	0.382**	1.155
	(0.028)	(0.129)
Observations	320	320

Notes: Dependent variable *withhold evidence* is equal to 1 if the low type agent sent no evidence in the Commitment treatment and 0 if they sent low evidence. *Period* takes values from 1 to 20 and represents the period. *Gender* is a dummy variable that takes the value 1 if subject is female and 0 otherwise. *Risk Aversion* takes the value 1 if the subject is classified as risk averse based on the number of safe options they chose in Activity 1 and 0 otherwise. *Ability to Bayesian update* is a dummy variable that takes the value 1 if subject answered the Activity 2 question of Part II correctly and 0 otherwise. p-values computed by score wild bootstrap procedure are in parentheses (clustered at the session level); * p<0.1, ** p<0.05, *** p<0.01.

Payoff of Subjects

Last, we turn our attention to the payoffs agents and principals. Experimental results show that the average payoff of agents is equal to 48.3 in the No-Commitment Treatment versus 59.8 in the Commitment Treatment. The difference is statistically significant ($p < 0.001$). A breakdown

of agents by type shows that both types earn significantly less in the No-Commitment Treatment. The payoff of low type agents is 46.1 in the No-Commitment Treatment versus 58.1 in the Commitment Treatment. Average payoff of the high type agents is 50.5 in the No-Commitment Treatment versus 61.7 in the Commitment Treatment. Both differences are statistically significant $(p < 0.001)$.

Result 5 *Both low and high types of agents have a higher payoff in the commitment treatment compared to the no-commitment treatment.*

Average payoff of the principal is 52.2 in the No-Commitment Treatment versus 51.3 in the Commitment Treatment. The difference is not statistically significant ($p = 0.756$). Note that we have exactly 50% split of high and low type agents in our experimental setting. If the agent type has a symmetric effect on principal payoff in opposing directions, the effect could be cancelling out. We also look at the average principal payoff conditional on agent type and see that this is indeed the case. Average payoff of the principal when the agent is low type is 53.9 in the No-Commitment Treatment versus 41.9 in the Commitment Treatment (difference is statistically significant, $p < 0.001$), and the average payoff of the principal when the agent is high type is 50.5 in the No-Commitment Treatment versus 61.7 in the Commitment Treatment (difference is statistically significant, $p < 0.001$).

Result 6 *When the agent is high type, the principal has a higher payoff in the Commitment Treatment compared to the No-Commitment Treatment. When the agent is low type, the principal has a higher payoff in the No-Commitment Treatment compared to the Commitment Treatment.*

3.5 Discussion

Even though our experimental setup satisfies all the conditions in Hart et al. (2017), we fail to find equivalence of outcomes between the No-Commitment and the Commitment treatments. When the outcomes of an evidence game with and without commitment do not coincide, there may be multiple possible reasons which can explain such a divergence in a more general setting. For example, in presence of multiplicity of equilibria, the difference between the outcomes could be caused by the equilibrium selection. Another explanation could be that subjects have trouble doing Bayesian updating (Friedman, 1998, Charness and Levin, 2005) or that senders strategically act truthfully to exploit receivers' naivete (Jin et al., 2021). However, our experimental design is intentionally simple enough to rule out these alternative explanations. Nevertheless, our experimental results yield that commitment leads to a difference in outcomes: even though the principals set rewards in line with equilibrium predictions when there is no-commitment, they consistently choose higher rewards when they commit on the reward scheme. Furthermore, despite the high reward for no-evidence, agents are still less likely to withhold their information when there is commitment. So, what accounts for our findings?

Our results highlight that when the agent is the second player, low type agents are less likely withhold their evidence even when it is profitable to do so. Hence, there should be an additional motive to payoff maximization. By allowing for lying costs in the framework of Hart et al. (2017), we show that our experimental findings are in line with the predictions of this costly lying model.[11]

[11] Alternatively, a low-type agent may feel guilty for disappointing the principal by withholding his evidence (see e.g. Battigalli and Dufwenberg, 2007; Charness and Dufwenberg, 2006). In Appendix C.3, we show that such a guilt aversion model does not predict our experimental findings accurately.

Model with Lying Averse Agents

Lying costs have been widely studied in a cheap talk setup since the seminal work of Kartik (2009). Our paper is the first to investigate it in an evidence game framework. In evidence games, the agents bear a cost of lying when they withhold evidence. Although agents need to lie to withhold their evidence in our experiment,[12] it is possible that the agents can hide the whole truth without lying in other setups. The experimental literature shows that subjects have preferences for truth-telling as well as lying-aversion, even though the strength of preferences might differ in magnitude (see e.g. Ertac et al., 2016; Friesen and Gangadharan, 2013; Sánchez-Pagés and Vorsatz, 2009; Serra-Garcia et al., 2011).

In evidence games, the relevance of lying costs has already been hinted in Hart et al. (2017). As a motivation for truth-leaning refinement, Hart et al. (2017) make a limit argument that the agent's utility increases infinitesimally if and only if when he does not withhold any evidence. Equivalently, the agent's utility decreases infinitesimally if and only if when he withholds an evidence.[13] We call this decrease in utility as cost of lying, and we allow it to be small but positive. Formally, simplifying Kartik (2009), we follow Serra-Garcia et al. (2013) such that the utility of an agent, with type t and cost of lying $k \geq 0$, sending a message m, and receiving a reward $x \geq 0$, $\hat{u}^A(m, x; t, k)$ takes the form:

$$\hat{u}^A(m, x; k, t) = \begin{cases} u(x) & \text{if truthful} \\ u(x) - k & \text{if withhold evidence} \end{cases}$$

[12] In order to withhold evidence, agents need to lie about not having low evidence (i.e. send "I don't have evidence for my type" even though they do have evidence).

[13] It is not uncommon to use preference for truth-telling and cost of lying interchangeably (see e.g. Abeler et al., 2019).

where $u(.)$ is continuously differentiable and strictly increasing.

On the other hand, since the principal does not send a message, her utility is assumed to be as in Section 2, $U^P(m, x; t) = w(x; v(t))$ where $w(.)$ is a continuously differentiable, strictly concave and single peaked function maximized when the reward is equal to the value of the agent, $x = v(t)$.

The agent can be one of two types: High or Low types with values $v(High) = H$ and $v(Low) = L$ such that $H > L \geq 0$. High type occurs with probability q, and Low type occurs with probability $1-q$ where $q \in (0, 1)$. Let $I > 0$ be the additional compensation to the principal. Denote $\rho = \frac{1-q}{q}$. Recall that for the parameters used in the experiment (I=100, H=100, L=0 and q=0.5), in the absence of lying costs (i.e. $k = 0$), $x_0^{NC} = 50$ and $x_-^{NC} = 0$ in the No-Commitment treatment, and $x_0^C = 50$ and $x_-^C \leq 50$ in the Commitment treatment.

Before characterizing the optimal mechanism of the Commitment setup and the equilibrium of the No-Commitment setup under lying averse agents, let's illustrate that lying averse agents might behave differently than what is predicted in the model without lying costs. Consider, for example, two policies that a principal can commit. Say, in both of these policies, the payoff of an agent who withholds his information is 50, and the payoff of an agent who reveals his information is 0 in Policy 1 but it is 49 in Policy 2. An agent, who does not have a lying cost ($k = 0$), withholds his information in both of the policies since $u(50) > u(0)$ and $u(50) > u(49)$. However, an agent, with a small but positive cost of lying such that $u(49) > u(50) - k > u(0)$, withholds his information in Policy 1 since $u(50)-k > u(0)$ but reveals his information in Policy 2 since $u(50) - k < u(49)$. Indeed, under lying aversion model, the outcomes of the equilibrium when there is no commitment and the optimal mechanism when there is commitment may not coincide.

Rewards for No Evidence and Low Evidence

When there is no commitment, in the unique sequential equilibrium, the principal sets $x_{-}^{NC} = L$ for any $k \geq 0$ since the low evidence could be provided only by the agent with low value. The agent with no evidence does not have any evidence to send, and the agent with low evidence sends no-evidence if k is small enough.[14] Then, the principal's problem when she sees no-evidence is:

$$\max_{x_0} \quad q \cdot w(I - |H - x_0|) + (1 - q) \cdot w(I - |L - x_0|) \tag{3.1}$$

The solution to this problem results in

$$w^{'}(I - H + x_0^{NC}) = \rho \cdot w^{'}(I + L - x_0^{NC}) \tag{3.2}$$

For the parameters of the experiment, Equation 3.2 becomes $w^{'}(x_0^{NC}) = w^{'}(100 - x_0^{NC})$, which implies that $x_0^{NC} = 50$. In other words, in the unique sequential equilibrium, the principal sets the reward for no evidence equal to 50. That is in line with Result 1(a).

When there is commitment, for any strictly positive cost of lying, $k > 0$, the optimal mechanism *can* separate the types. Recall that in the absence of cost of lying, in order to separate the types, the principal needs to give distinct rewards, i.e. $x_{-} \neq x_0$. Also, the reward for low evidence needs to be higher than the reward for no evidence, i.e. $x_{-} > x_0$, otherwise, i.e. $x_{-} < x_0$, the low type agent will withhold his low evidence since $u(x_{-}) < u(x_0)$. But this cannot

[14] Note that there should be an upper bound on the cost of lying, since if k were very large, the rewards would have become irrelevant for the subjects, and they would always reveal their evidence no matter what the rewards are. Such behavior is not observed in our data. We will show that this additional complications are not necessary, and our data can be explained by small cost of lying.

be optimal since the value of the low type agent is smaller than the value of the high type agent. On the other hand, in the presence of cost of lying, the reward for low evidence can be lower than the reward for no evidence, i.e. $x_- < x_0$, and the low type agent may still reveal his low evidence since $u(x_-) > u(x_0) - k$. So, for $k > 0$, the principal's problem is

$$\begin{aligned} \max_{x_0, x_-} \quad & q \cdot w(I - |H - x_0|) + (1 - q) \cdot w(I - |L - x_-|) \\ s.t. \quad & u(x_-) \geq u(x_0) - k \\ & u(x_0) \geq u(x_-) \geq 0 \end{aligned} \tag{3.3}$$

Then in the optimal mechanism

$$\begin{aligned} u(x_-^C) &= u(x_0^C) - k, \text{ and} \\ w^{'}(I - H + x_0^C) &= \rho \cdot w^{'}(I + L - x_-^C) \end{aligned} \tag{3.4}$$

For the parameters of the experiment, Equation 3.4 becomes $w^{'}(x_0^C) = w^{'}(100 - x_-^C)$. It implies that for a concave $w(.)$, $x_-^C + x_0^C = 100$. So, $0 < x_-^C < 50 < x_0^C$ which is in line with Result 1(b).

Additionally, in Proposition 1 we show that when the cost of lying is strictly positive, the commitment matters such that the principal sets a higher reward for no evidence when there is a commitment than when there is no commitment (which is in line with Result 2).

Proposition 1 $x_0^C > x_0^{NC}$.

Proof: The only important assumption regarding $u(.)$ that it is a strictly increasing function. So, w.l.o.g., let $u(x) = x$. Equation 3.2 remains the same, and Equation 3.4 becomes:

$$x_{-}^{C} = x_0^{C} - k, \text{ and}$$
$$w^{'}(I - H + x_0^{C}) = \rho \cdot w^{'}(I + L - x_0^{C} + k) \tag{3.5}$$

For contradiction, assume $x_0^{C} \leq x_0^{NC}$. Then, for any $k > 0$, $-x_0^{C} + k > -x_0^{NC}$. So,

$$w^{'}(I - H + x_0^{C}) = \rho \cdot w^{'}(I + L - x_0^{C} + k) < \rho \cdot w^{'}(I + L - x_0^{NC}) = w^{'}(I - H + x_0^{NC}) \tag{3.6}$$

where the first and the last equalities follow from Equations 3.2 and 3.5, the inequality follows from strict concavity of $w(.)$. But $w^{'}(I - H + x_0^{C}) < w^{'}(I - H + x_0^{NC})$ implies that $x_0^{C} > x_0^{NC}$ which is a contradiction. ■

Withholding Evidence

Next, we look at the effect of rewards on subjects' decision to withhold their evidence. A low type agent withholds his evidence if $u(x_{-}) < u(x_0) - k$. As argued above, in the No-Commitment treatment, every low type agent with a small cost of lying withholds his evidence since $u(0) < u(50) - k$. However, the agent with the same cost of lying may reveal his evidence in the Commitment treatment, since the optimal mechanism can incentivize not withholding the evidence by setting rewards such that $u(x_{-}^{C}) = u(x_0^{C}) - k$. To see this, for example, consider a low type agent with $u(x) = x$ and $k = 20$. Say, a principal commits to the reward 60 if the agent provides no evidence and the reward 40 if the agent reveals low evidence. A low type agent with $u(x) = x$ and $k = 20$ will not withhold his evidence since $40 = 60 - 20$, but such an agent will withhold his evidence in the No-Commitment treatment since $0 < 50 - 20$. Hence, there will

be fewer low type agents withholding their evidence in the Commitment treatment (in line with Result 3).

Additionally, in the Commitment treatment, consider the cases where the reward for no evidence, x_0^C is higher than the reward for low evidence, x_-^C. Still, any low type agent with $k > 0$ reveals his evidence if and only if $u(x_-^C) \geq u(x_0^C) - k$. Since by changing the rewards it is possible to change the direction of the inequality, the decision of the agent may be altered. In particular, for any low type agent with $k > 0$, there is a positive relation between the reward for no evidence and the likelihood of withholding the evidence, and a negative relation between the reward for low evidence and the likelihood of withholding the evidence (in line with Result 4). To see this, suppose $u(x_-^C) \geq u(x_0^C) - k$, i.e. agent reveals his evidence. If the reward for no evidence decreases to $\hat{x}_-^C$ such that $u(\hat{x}_-^C) < u(x_0^C) - k$, he withholds his evidence; or if the reward for low evidence increases to $\tilde{x}_0^C$ such that $u(x_-^C) < u(\tilde{x}_0^C) - k$, he withholds his evidence. Similarly, suppose $u(x_-^C) < u(x_0^C) - k$, i.e. he withholds his evidence. If the reward for no evidence increases to $\tilde{x}_-^C$ such that $u(\tilde{x}_-^C) \geq u(x_0^C) - k$, he reveals his evidence; or if the reward for low evidence decreases to $\hat{x}_0^C$ such that $u(x_-^C) \geq u(\hat{x}_0^C) - k$, he reveals his evidence.

Welfare Implications

Finally, if the outcome equivalence between two setups does not hold, does the principal prefer to commit on a policy when she faces a lying averse agent? We have shown that when the agent is lying averse, the principal sets higher rewards both for low evidence and no evidence in a committed policy than in no-commitment. On the other hand, the principal can separate the types only with commitment. In this trade-off, it turns out to be that principal is better off in a

committed policy.

Proposition 2 *For $k > 0$, principal's expected utility when there is commitment is higher than that when there is no commitment.*

Proof: Since $x_0^C > x_0^{NC}$, $w^{'}(I - H + x_0^C) < w^{'}(I - H + x_0^{NC})$ due to strict concavity of $w(.)$. Plugging in the corresponding expressions from Equations 3.2 and 3.5, we get $\rho \cdot w^{'}(I + L - x_0^C + k) < \rho \cdot w^{'}(I + L - x_0^{NC})$ which in turn results in $x_0^C - k < x_0^{NC}$ by strict concavity of $w(.)$.

Principal's expected utility when there is commitment is

$$
\begin{aligned}
& q \cdot w(I - H + x_0^C) + (1 - q) \cdot w(I + L - x_-^C) \\
& = q \cdot w(I - H + x_0^C) + (1 - q) \cdot w(I + L - x_0^C + k) \text{ (since } x_-^C = x_0^C - k) \\
& > q \cdot w(I - H + x_0^{NC}) + (1 - q) \cdot w(I + L - x_0^C + k) \text{ (since } x_0^C > x_0^{NC} \text{by Proposition 1)} \\
& > q \cdot w(I - H + x_0^{NC}) + (1 - q) \cdot w(I + L - x_0^{NC}) \text{ (since } x_0^{NC} > x_0^C - k)
\end{aligned}
\tag{3.7}
$$

that is Principal's expected utility when there is no commitment. Hence, principal is better off in a setup with commitment. ■

For example, for the parameters of the experiment, with lying averse agents, in the unique equilibrium without commitment, $x_0^{NC} = 50$ and $x_-^{NC} = 0$. So, when the principal does not commit on a policy, her expected utility is $0.5 * w(100 - (100 - 50) + 0.5 * w(100 - (50 - 0)) = w(50)$. When there is commitment, the optimal mechanism separates the types with the rewards such that $x_0^C > 50 > x_-^C$ and $x_0^C + x_-^C = 100$. So, her expected utility with commitment is $0.5 * w(100 - (100 - x_0^C)) + 0.5 * w(100 - x_-^C) = w(x_0^C) > w(50)$. Hence, when the agent is lying averse, the principal prefers to have a committed policy.

3.6 Conclusion

The role of communication has been widely investigated as a form of cheap-talk and Bayesian persuasion games (see e.g. Fréchette et al., 2019 for an experimental analysis of subject behavior when senders have commitment power). Our experiment is the first to test whether commitment has any value in evidence games, in which an uninformed receiver who chooses a reward for an informed sender who can reveal pieces of evidence about his type. In a setup without commitment, the receiver moves second and chooses a reward after observing the sender's message; while in a setup with commitment, the receiver commits to a reward scheme in advance and the sender chooses which message to send after observing the rewards.

We design our experiment simple enough (a simpler version of Example 1 of Hart et al., 2017) to leave minimum room for subject mistakes. We experimentally falsify the equivalence of outcomes between these two setups, contrary to the theoretical predictions based on Hart et al. (2017). Our experimental results yield that although subjects behave in line with the equilibrium predictions when there is no commitment, they consistently choose higher rewards when they commit on the reward scheme. Hence, commitment has value. We show that the predictions of a model that includes cost of lying to the standard model are in line with our experimental findings. Additionally, when facing with a lying averse agent, we theoretically demonstrate that the principal is better off from a committed policy. It may be interesting to experimentally investigate this theoretical prediction. Particularly, if a principal is given an option to decide whether to commit on a policy or not, is she willing to pay to commit on a policy? We leave this question for future work.

Appendix A: Appendix to Chapter 1

A.1 Additional Tables and Figures

Table A.1: Demographics Breakdown Across Treatments

	Treatment			p-values		
	N	NC	C	N-NC	N-C	NC-C
Gender						
Male	50.5%	50.3%	50.2%	0.97	0.93	0.97
Female	49.5%	49.7%	49.8%	0.97	0.93	0.97
Age	37.21	35.70	35.27	0.31	0.13	0.59
Education						
High School or Less	13.90	11.49	7.02	0.38	0.01	0.06
Associate Degree	9.15	11.15	10.37	0.42	0.62	0.76
Some College	23.39	23.65	27.76	0.94	0.22	0.25
Bachelor's Degree	37.29	37.16	39.80	0.97	0.53	0.51
Post Graduate Degree	16.27	16.55	15.05	0.93	0.68	0.62
Income						
Less than $20K	10.51	13.51	12.04	0.26	0.56	0.59
Between $20K and $30K	11.86	11.15	9.70	0.79	0.39	0.56
Between $30K and $50K	18.31	19.59	16.72	0.69	0.61	0.36
Between $50K and $70K	19.66	18.92	21.40	0.82	0.60	0.45
Between $70K and $150K	29.49	24.32	26.42	0.16	0.40	0.56
More than $150K	10.17	12.50	13.71	0.37	0.18	0.66
N	295	296	299			

Notes: The columns N, NC, and C correspond to *Noisy*, *NoisyComparative*, and *Comparative* Treatments, respectively. The last three columns compare values across the associated treatment pairs and report p-values obtained by a test of proportions (for ratios) or by a Mann Whitney U-test (for the continuous variable age).

Table A.2: Number of Subjects in Each Treatment Broken Down by Gender and Signal Received

	Bad News			Good News		
	N	NC	C	N	NC	C
$Female$	68	54	59	60	68	73
$Male$	59	59	65	74	74	70

Notes: There is no significant difference in percent of males and females who receive good news in any treatment. p-values obtained by a test of proportions are 0.157, 0.987, and 0.572 for *Noisy*, *NoisyComparative*, and *Comparative* Treatments, respectively.

Table A.3: OLS Regressions Relating Posterior Beliefs on Gender and Individual Characteristics

	Bad News			Good News		
Coefficient	N	NC	C	N	NC	C
$Female$	-0.545***	-0.127	-0.140	-0.019	-0.108	-0.385**
	(0.199)	(0.266)	(0.202)	(0.186)	(0.196)	(0.159)
$Prior$	0.658***	0.760***	0.740***	0.629***	0.672***	0.602***
	(0.051)	(0.067)	(0.048)	(0.043)	(0.055)	(0.037)
$Score$	0.078**	0.004	0.062	-0.029	0.009	0.039
	(0.030)	(0.044)	(0.044)	(0.025)	(0.028)	(0.025)
$Education$	-0.117	0.121	-0.231	0.124	-0.072	-0.109
	(0.211)	(0.246)	(0.204)	(0.205)	(0.230)	(0.158)
$Income$	0.128	-0.078	0.089	-0.036	0.239	0.149
	(0.202)	(0.260)	(0.207)	(0.203)	(0.217)	(0.155)
Age	-0.007	-0.008	-0.002	-0.001	-0.016*	0.005
	(0.006)	(0.010)	(0.008)	(0.008)	(0.009)	(0.007)
$Constant$	-1.103***	-0.564	-0.983**	1.434***	1.431***	0.421
	(0.399)	(0.554)	(0.468)	(0.439)	(0.432)	(0.367)
N	127	113	124	134	142	143
R^2	0.704	0.617	0.712	0.676	0.626	0.730

Notes: Dependent variable is the is the posterior log-likelihood ratio for being among top-half performers, $Female$ is a dummy variable equal to 1 if the gender is female and 0 if male, $Prior$ is the prior log-likelihood ratio for being among top-half performers, $Score$ is the number of correct answers in the IQ test, $Education$ is a dummy variable equal to 1 if education is Bachelors' Degree or higher and 0 otherwise. $Income$ is a dummy variable equal to 1 if annual income is higher than $\$50k$ and 0 otherwise. Standard errors are reported in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

Table A.4: OLS Regressions Relating Posterior Beliefs on Gender and Individual Characteristics Without Controlling for Prior Beliefs

	Bad News			Good News		
Coefficient	N	NC	C	N	NC	C
$Female$	-1.174***	-1.241***	-0.656*	-0.370	-0.634**	-0.990***
	(0.297)	(0.368)	(0.344)	(0.299)	(0.276)	(0.262)
$Score$	0.211***	0.047	0.277***	0.102***	0.122***	0.158***
	(0.044)	(0.065)	(0.072)	(0.038)	(0.038)	(0.040)
$Education$	0.402	0.161	-0.140	0.680**	0.784**	0.267
	(0.317)	(0.366)	(0.353)	(0.327)	(0.317)	(0.265)
$Income$	-0.192	0.474	0.204	0.322	0.282	0.095
	(0.308)	(0.380)	(0.356)	(0.327)	(0.314)	(0.264)
Age	-0.009	-0.005	-0.002	0.006	-0.032**	0.007
	(0.010)	(0.014)	(0.013)	(0.013)	(0.013)	(0.012)
$Constant$	-1.721***	-0.668	-2.109***	-0.489	0.974	-0.674
	(0.609)	(0.824)	(0.797)	(0.679)	(0.622)	(0.614)
N	127	113	124	134	142	143
R^2	0.296	0.146	0.137	0.141	0.212	0.213

Notes: Dependent variable is the is the posterior log-likelihood ratio for being among top-half performers, $Female$ is a dummy variable equal to 1 if the gender is female and 0 if male, $Score$ is the number of correct answers in the IQ test, $Education$ is a dummy variable equal to 1 if education is Bachelors' Degree or higher and 0 otherwise. $Income$ is a dummy variable equal to 1 if annual income is higher than $\$50k$ and 0 otherwise. Standard errors are reported in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

A.2 Truncation of Extreme Beliefs

For prior beliefs over ranks, I consider 6 different truncation methods. I use method 1 for the results reported in the main body of the paper. All results are robust to using the other 5 methods instead. In all below equations, p_i denotes the prior belief on rank $i \in \{1, 2, ..., 10\}$ before truncation, $Pr(H)$ denotes the prior belief of being among top-half performers before truncation, $Pr(L)$ denotes the prior belief of being among bottom-half performers before truncation, p_i^* denotes the same belief after truncation, n_0 denotes the number of ranks with 0 prior belief, n_0^{top} denotes the number of ranks in the top-half with 0 prior belief, and n_0^{bottom} denotes the number of ranks in the bottom-half with 0 prior belief.

Method 1

This method replaces probabilities equal to 0 with 0.2 and subtracts the total added probability from all non-zero probability ranks weighted by the prior in the corresponding rank.

$$\text{Replace } p_i^* = 0.2 \text{ for all ranks } i \text{ with } p_i = 0$$

$$\text{Replace } p_i^* = p_i - \frac{p_i \times 0.2 \times n_0}{100} \text{ for all ranks } i \text{ with } p_i \neq 0$$

Method 2

This method replaces probabilities equal to 0 with $1/n_0$ and subtracts a total of 1% probability from all non-zero probability ranks weighted by the prior in the corresponding rank.

$$\text{Replace } p_i^* = \frac{1}{n_0} \text{ for all ranks } i \text{ with } p_i = 0$$

$$\text{Replace } p_i^* = p_i - \frac{p_i}{100} \text{ for all ranks } i \text{ with } p_i \neq 0$$

Method 3

This method is similar to Method 1, with the addition that it does not modify the probability of being among top-half or bottom-half performers for a subject. It replaces probabilities equal to 0 with 0.2. If $Pr(H)$ and $Pr(L)$ are both strictly positive, it subtracts the total added probability to each half from all non-zero probability ranks in that half weighted by the prior. If either $Pr(H)$ or $Pr(L)$ is equal to 0, then this method is identical to Method 1.

$$\text{Replace } p_i^* = 0.2 \text{ for all ranks } i \text{ with } p_i = 0$$

Case 1: *Pr(H)>0 and Pr(L)>0*

$$\text{Replace } p_i^* = p_i - \frac{p_i \times 0.2 \times n_0^{top}}{Pr(H)} \text{ for all ranks in the top half } i \text{ with } p_i \neq 0$$

$$\text{Replace } p_i^* = p_i - \frac{p_i \times 0.2 \times n_0^{bottom}}{Pr(L)} \text{ for all ranks in the bottom half } i \text{ with } p_i \neq 0$$

Case 2: *Pr(H)=0*

$$\text{Replace } p_i^* = p_i - \frac{p_i \times 0.2 \times n_0}{100} \text{ for all ranks in the bottom half } i \text{ with } p_i \neq 0$$

Case 3: *Pr(L)=0*

$$\text{Replace } p_i^* = p_i - \frac{p_i \times 0.2 \times n_0}{100} \text{ for all ranks in the top half } i \text{ with } p_i \neq 0$$

Method 4

This method is similar to Method 2, with the addition that it does not modify the probability of being among top-half or bottom-half performers for a subject. It replaces probabilities equal to 0 with $1/n_0$. If $Pr(H)$ and $Pr(L)$ are both strictly positive, it subtracts the total added probability to each half from all non-zero probability ranks in that half weighted by the prior. If either $Pr(H)$ or $Pr(L)$ is equal to 0, then this method is identical to Method 2.

$$\text{Replace } p_i^* = \frac{1}{n_0} \text{ for all ranks } i \text{ with } p_i = 0$$

Case 1: *Pr(H)>0 and Pr(L)>0*

$$\text{Replace } p_i^* = p_i - \frac{p_i \times (n_0^{top}/n_0)}{Pr(H)} \text{ for all ranks in the top half } i \text{ with } p_i \neq 0 \text{ if } n_0 \neq 0$$

$$\text{Replace } p_i^* = p_i - \frac{p_i \times (n_0^{bottom}/n_0)}{Pr(L)} \text{ for all ranks in the bottom half } i \text{ with } p_i \neq 0 \text{ if } n_0 \neq 0$$

Case 2: *Pr(H)=0*

$$\text{Replace } p_i^* = p_i - \frac{p_i}{100} \text{ for all ranks in the bottom half } i \text{ with } p_i \neq 0$$

Case 3: *Pr(L)=0*

$$\text{Replace } p_i^* = p_i - \frac{p_i}{100} \text{ for all ranks in the top half } i \text{ with } p_i \neq 0$$

Method 5

This method replaces the zero-probabilities in a given half with 0.2 only if all rank probabilities in that half is 0, so that a total of 1% is subtracted from the half that has 0 probability. It subtracts a total of 1% probability from all non-zero probability ranks in the other half weighted by the prior in the corresponding rank.

Case 1: *Pr(H)=0*

Replace $p_i^* = 0.2$ for all ranks in the top half i with $p_i = 0$

Replace $p_i^* = p_i - \frac{p_i}{100}$ for all ranks in the bottom half i with $p_i \neq 0$

Case 2: *Pr(L)=0*

Replace $p_i^* = 0.2$ for all ranks in the bottom half i with $p_i = 0$

Replace $p_i^* = p_i - \frac{p_i}{100}$ for all ranks in the top half i with $p_i \neq 0$

Method 6

This method replaces the zero-probabilities in a given half with a total of 1% weighted by the number of zero-probability ranks in the corresponding rank. It subtracts a total of 1% probability from each half equally split between all non-zero probability ranks in the relevant half.

Replace $p_i^* = \frac{1}{n_0^{top}}$ for all ranks in the top half i with $p_i = 0$

Replace $p_i^* = \frac{1}{n_0^{bottom}}$ for all ranks in the bottom half i with $p_i = 0$

Case 1: *Pr(H)>0 and Pr(L)>0*

Replace $p_i^* = p_i - \frac{1}{5 - n_0^{top}}$ for all ranks in the top half i with $p_i \neq 0$ if $n_0^{top} \neq 0$

Replace $p_i^* = p_i - \frac{1}{5 - n_0^{bottom}}$ for all ranks in the bottom half i with $p_i \neq 0$ if $n_0^{bottom} \neq 0$

Case 2: *Pr(H)=0*

Replace $p_i^* = p_i - 0.2$ for all ranks in the bottom half i with $p_i \neq 0$

Case 3: *Pr(L)=0*

Replace $p_i^* = p_i - 0.2$ for all ranks in the top half i with $p_i \neq 0$

A.3 Instructions

Figure A.1: Welcome Page

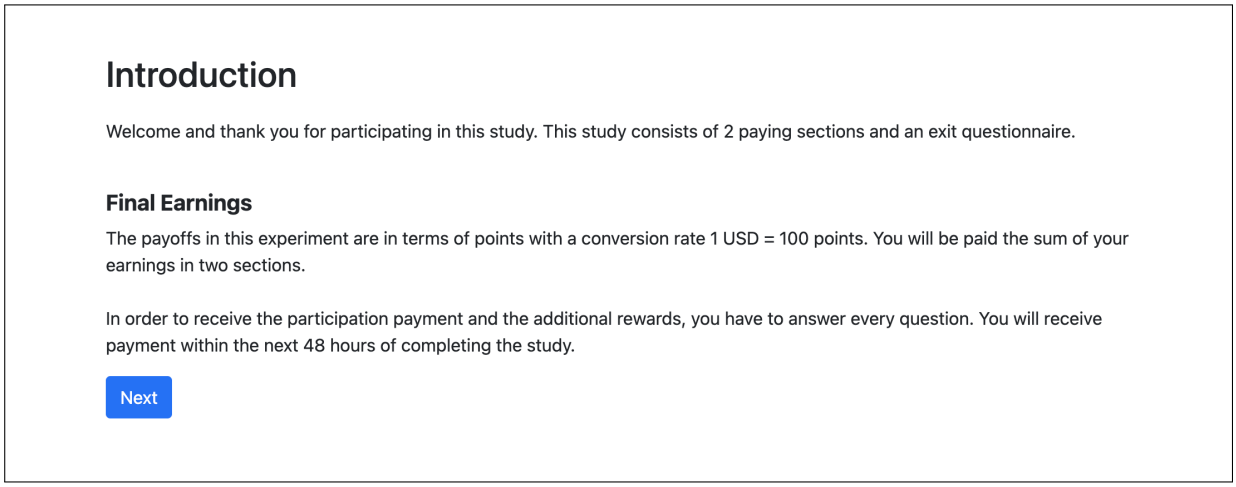


Figure A.2: Part I, Introduction

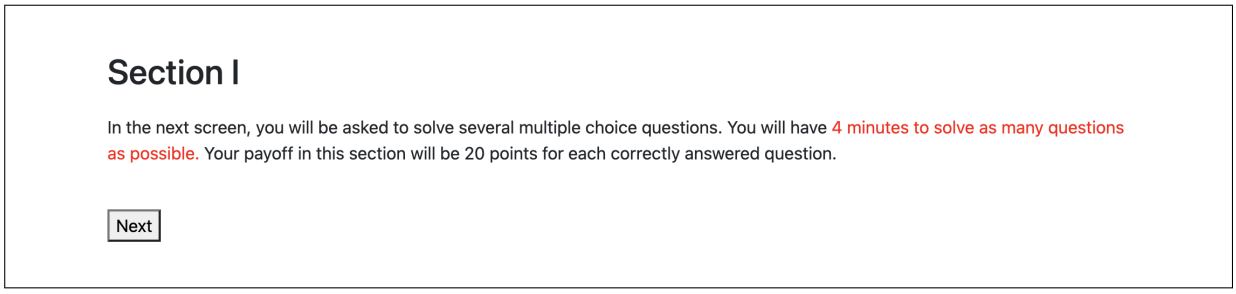


Figure A.3: Part II, Introduction

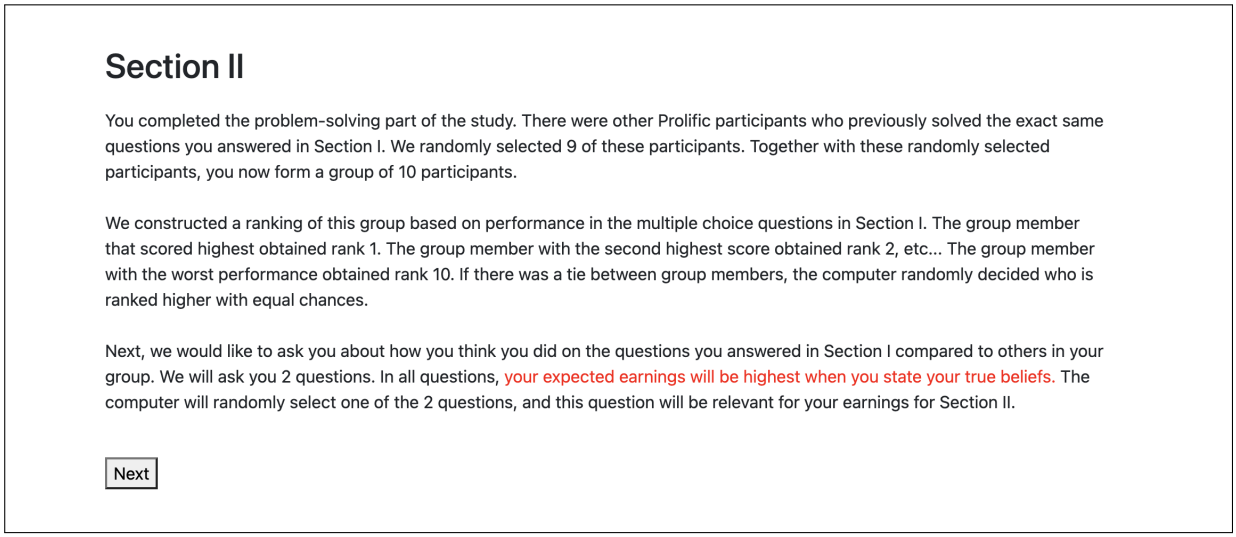
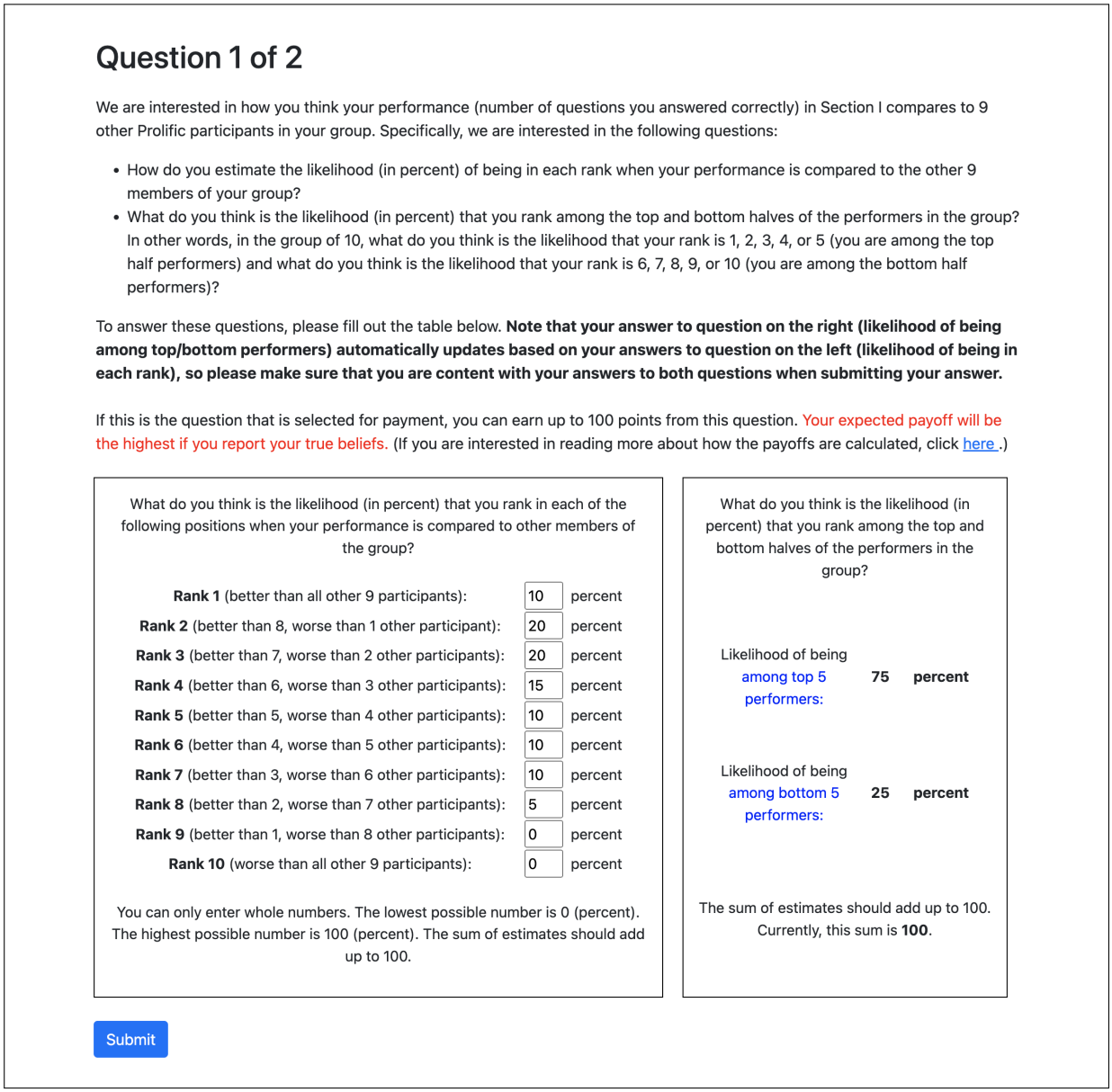


Figure A.4: Part II, Prior Belief Elicitation



Question 1 of 2

We are interested in how you think your performance (number of questions you answered correctly) in Section I compares to 9 other Prolific participants in your group. Specifically, we are interested in the following questions:

- How do you estimate the likelihood (in percent) of being in each rank when your performance is compared to the other 9 members of your group?
- What do you think is the likelihood (in percent) that you rank among the top and bottom halves of the performers in the group? In other words, in the group of 10, what do you think is the likelihood that your rank is 1, 2, 3, 4, or 5 (you are among the top half performers) and what do you think is the likelihood that your rank is 6, 7, 8, 9, or 10 (you are among the bottom half performers)?

To answer these questions, please fill out the table below. **Note that your answer to question on the right (likelihood of being among top/bottom performers) automatically updates based on your answers to question on the left (likelihood of being in each rank), so please make sure that you are content with your answers to both questions when submitting your answer.**

If this is the question that is selected for payment, you can earn up to 100 points from this question. Your expected payoff will be the highest if you report your true beliefs. (If you are interested in reading more about how the payoffs are calculated, click here .)

What do you think is the likelihood (in percent) that you rank in each of the following positions when your performance is compared to other members of the group?

Rank 1 (better than all other 9 participants):	10	percent
Rank 2 (better than 8, worse than 1 other participant):	20	percent
Rank 3 (better than 7, worse than 2 other participants):	20	percent
Rank 4 (better than 6, worse than 3 other participants):	15	percent
Rank 5 (better than 5, worse than 4 other participants):	10	percent
Rank 6 (better than 4, worse than 5 other participants):	10	percent
Rank 7 (better than 3, worse than 6 other participants):	10	percent
Rank 8 (better than 2, worse than 7 other participants):	5	percent
Rank 9 (better than 1, worse than 8 other participants):	0	percent
Rank 10 (worse than all other 9 participants):	0	percent

You can only enter whole numbers. The lowest possible number is 0 (percent). The highest possible number is 100 (percent). The sum of estimates should add up to 100.

What do you think is the likelihood (in percent) that you rank among the top and bottom halves of the performers in the group?

Likelihood of being among top 5 performers: **75** **percent**

Likelihood of being among bottom 5 performers: **25** **percent**

The sum of estimates should add up to 100. Currently, this sum is **100**.

Submit

Figure A.5: Part II, Prior Beliefs Pop-Up Box

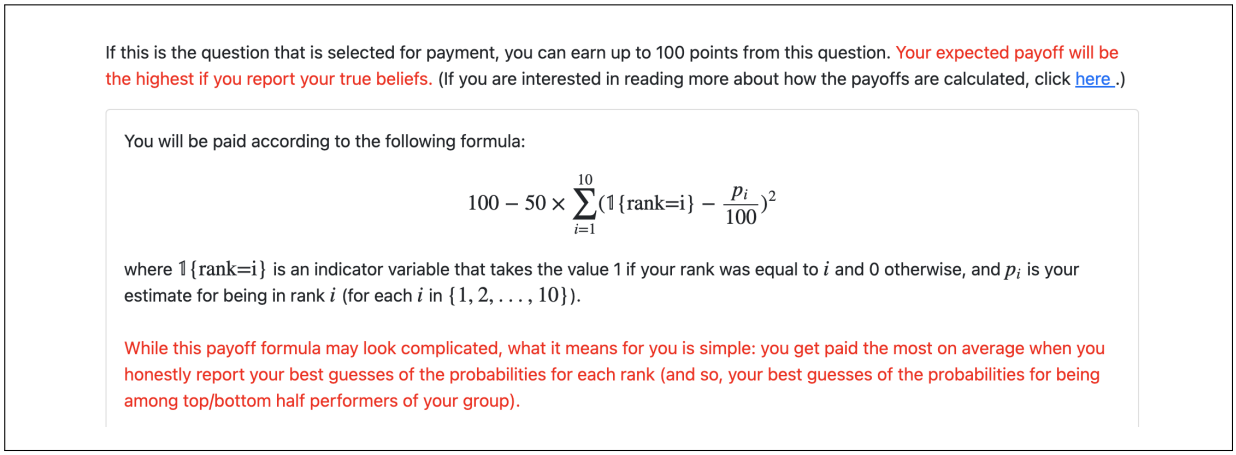


Figure A.6: Part II, Prior Belief Confirmation

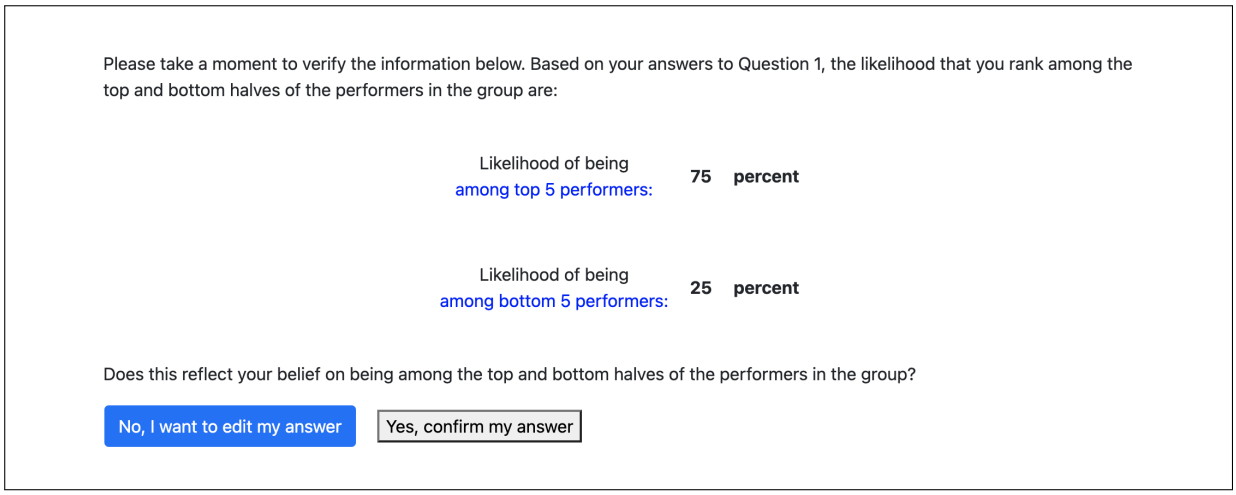
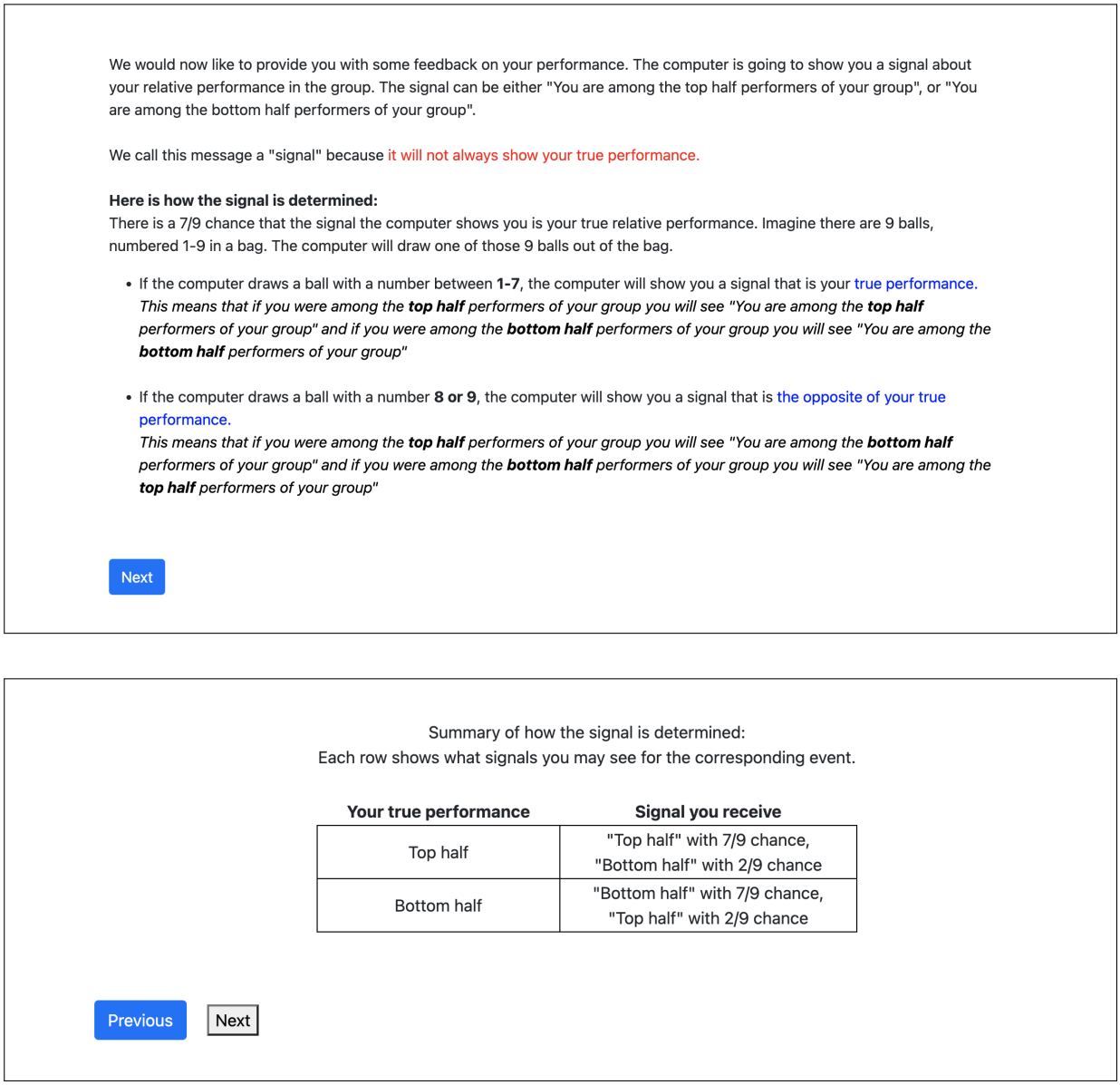


Figure A.7: Part II, Signal Instructions, Noisy Signal



We would now like to provide you with some feedback on your performance. The computer is going to show you a signal about your relative performance in the group. The signal can be either "You are among the top half performers of your group", or "You are among the bottom half performers of your group".

We call this message a "signal" because it will not always show your true performance.

Here is how the signal is determined:
There is a 7/9 chance that the signal the computer shows you is your true relative performance. Imagine there are 9 balls, numbered 1-9 in a bag. The computer will draw one of those 9 balls out of the bag.

- If the computer draws a ball with a number between **1-7**, the computer will show you a signal that is your true performance.
  *This means that if you were among the **top half** performers of your group you will see "You are among the **top half** performers of your group" and if you were among the **bottom half** performers of your group you will see "You are among the **bottom half** performers of your group"*

- If the computer draws a ball with a number **8 or 9**, the computer will show you a signal that is the opposite of your true performance.
  *This means that if you were among the **top half** performers of your group you will see "You are among the **bottom half** performers of your group" and if you were among the **bottom half** performers of your group you will see "You are among the **top half** performers of your group"*

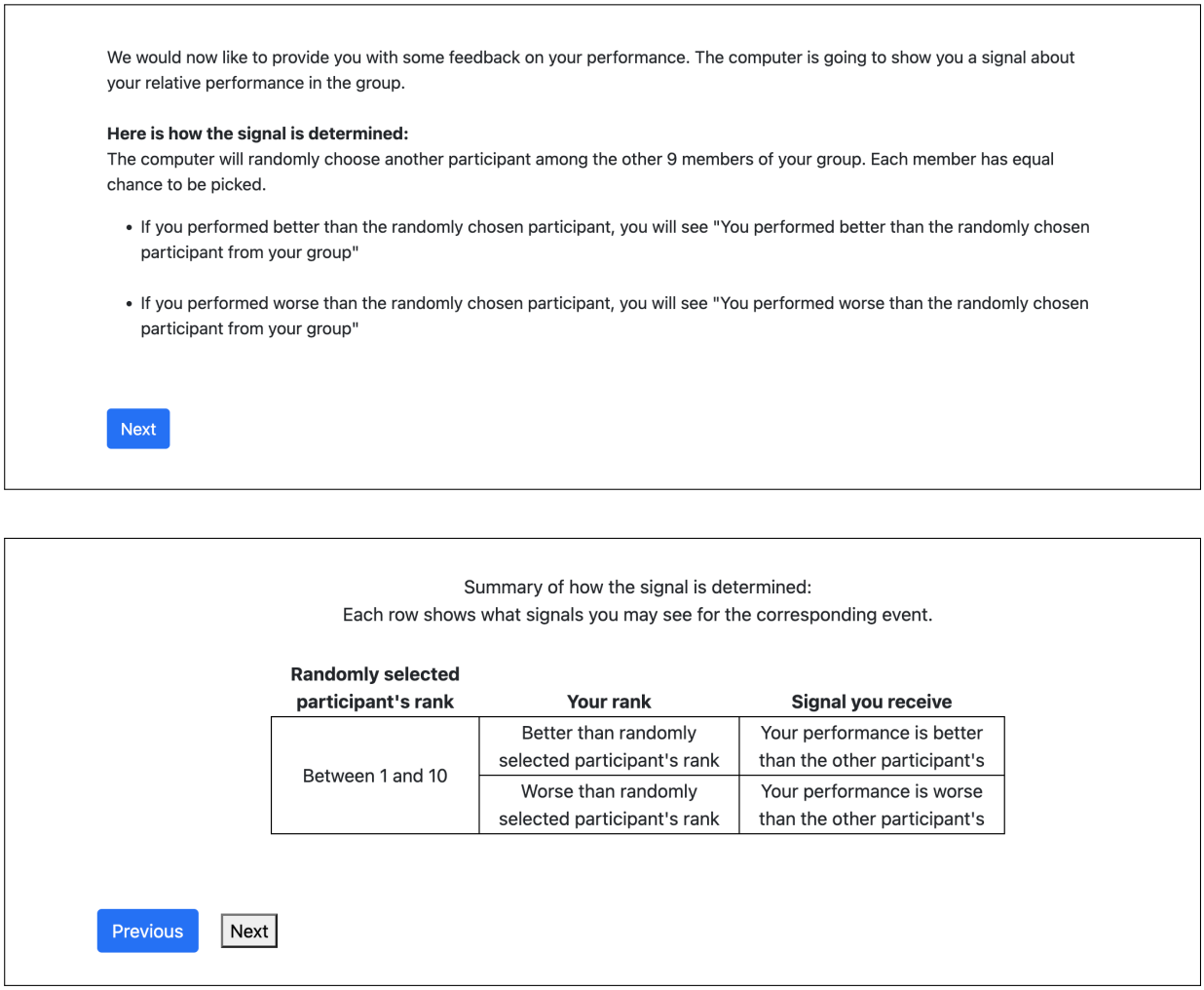
Next

Summary of how the signal is determined:
Each row shows what signals you may see for the corresponding event.

Your true performance	Signal you receive
Top half	"Top half" with 7/9 chance, "Bottom half" with 2/9 chance
Bottom half	"Bottom half" with 7/9 chance, "Top half" with 2/9 chance

Previous Next

Figure A.8: Part II, Signal Instructions, Comparative Signal



We would now like to provide you with some feedback on your performance. The computer is going to show you a signal about your relative performance in the group.

Here is how the signal is determined:
The computer will randomly choose another participant among the other 9 members of your group. Each member has equal chance to be picked.

- If you performed better than the randomly chosen participant, you will see "You performed better than the randomly chosen participant from your group"
- If you performed worse than the randomly chosen participant, you will see "You performed worse than the randomly chosen participant from your group"

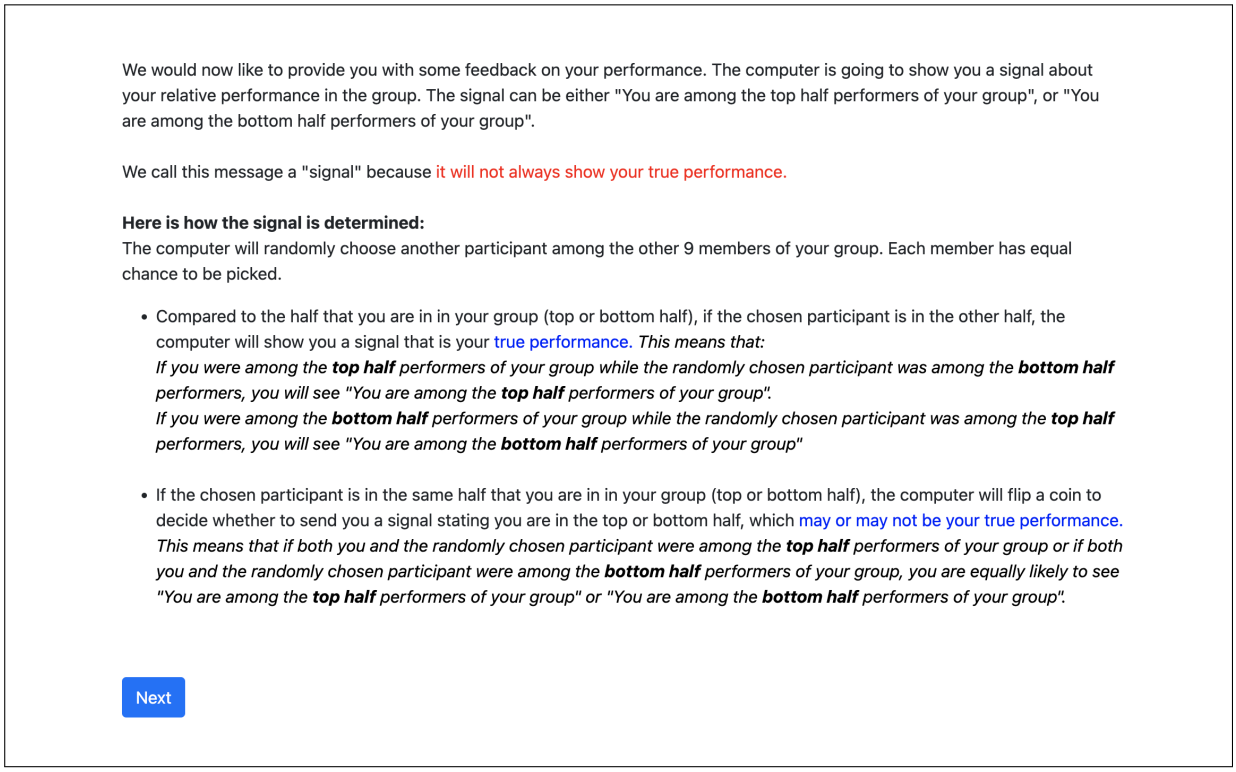
Next

Summary of how the signal is determined:
Each row shows what signals you may see for the corresponding event.

Randomly selected participant's rank	Your rank	Signal you receive
Between 1 and 10	Better than randomly selected participant's rank	Your performance is better than the other participant's
	Worse than randomly selected participant's rank	Your performance is worse than the other participant's

Previous Next

Figure A.9: Part II, Signal Instructions, NoisyComparative Signal



We would now like to provide you with some feedback on your performance. The computer is going to show you a signal about your relative performance in the group. The signal can be either "You are among the top half performers of your group", or "You are among the bottom half performers of your group".

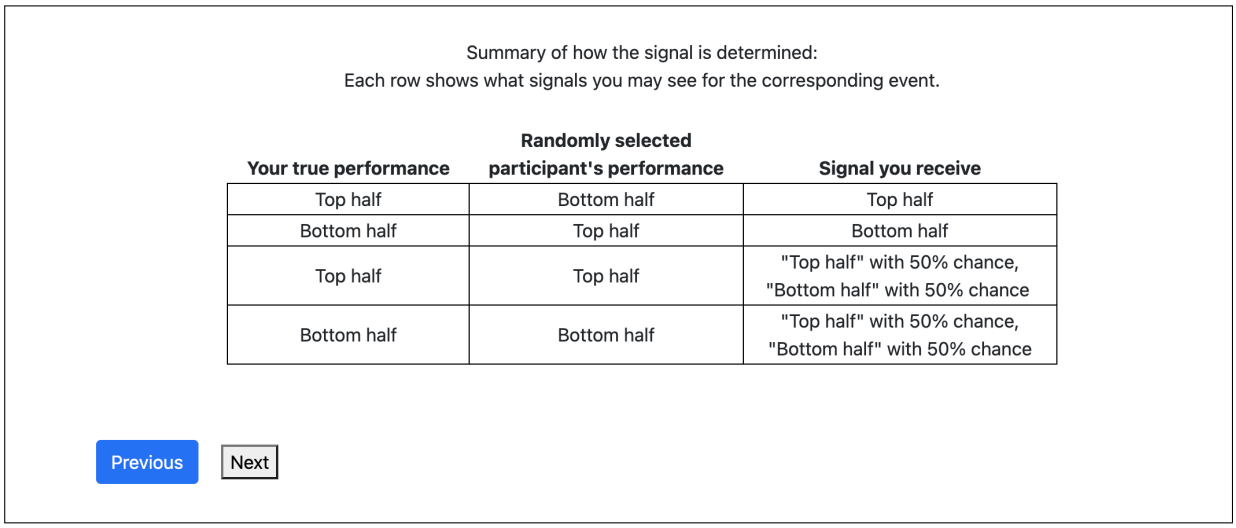
We call this message a "signal" because it will not always show your true performance.

Here is how the signal is determined:
The computer will randomly choose another participant among the other 9 members of your group. Each member has equal chance to be picked.

- Compared to the half that you are in in your group (top or bottom half), if the chosen participant is in the other half, the computer will show you a signal that is your true performance. *This means that:*
  *If you were among the **top half** performers of your group while the randomly chosen participant was among the **bottom half** performers, you will see "You are among the **top half** performers of your group".*
  *If you were among the **bottom half** performers of your group while the randomly chosen participant was among the **top half** performers, you will see "You are among the **bottom half** performers of your group"*

- If the chosen participant is in the same half that you are in in your group (top or bottom half), the computer will flip a coin to decide whether to send you a signal stating you are in the top or bottom half, which may or may not be your true performance. *This means that if both you and the randomly chosen participant were among the **top half** performers of your group or if both you and the randomly chosen participant were among the **bottom half** performers of your group, you are equally likely to see "You are among the **top half** performers of your group" or "You are among the **bottom half** performers of your group".*

Next



Summary of how the signal is determined:
Each row shows what signals you may see for the corresponding event.

Your true performance	Randomly selected participant's performance	Signal you receive
Top half	Bottom half	Top half
Bottom half	Top half	Bottom half
Top half	Top half	"Top half" with 50% chance, "Bottom half" with 50% chance
Bottom half	Bottom half	"Top half" with 50% chance, "Bottom half" with 50% chance

Previous Next

Figure A.10: Part II, Comprehension Question, Noisy Signal

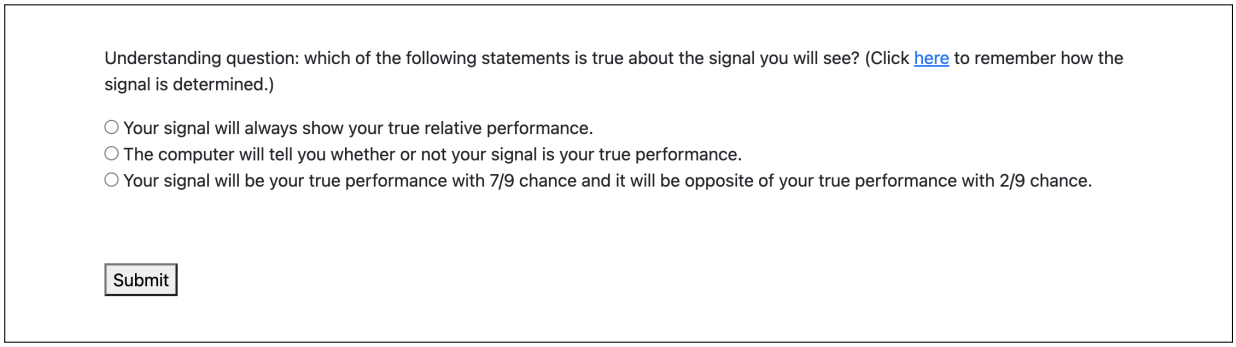


Figure A.11: Part II, Comprehension Question, Comparative Signal

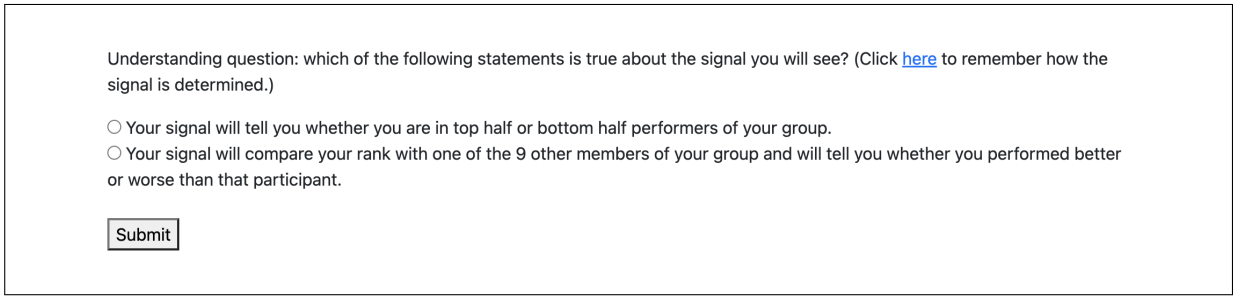


Figure A.12: Part II, Comprehension Question, NoisyComparative Signal

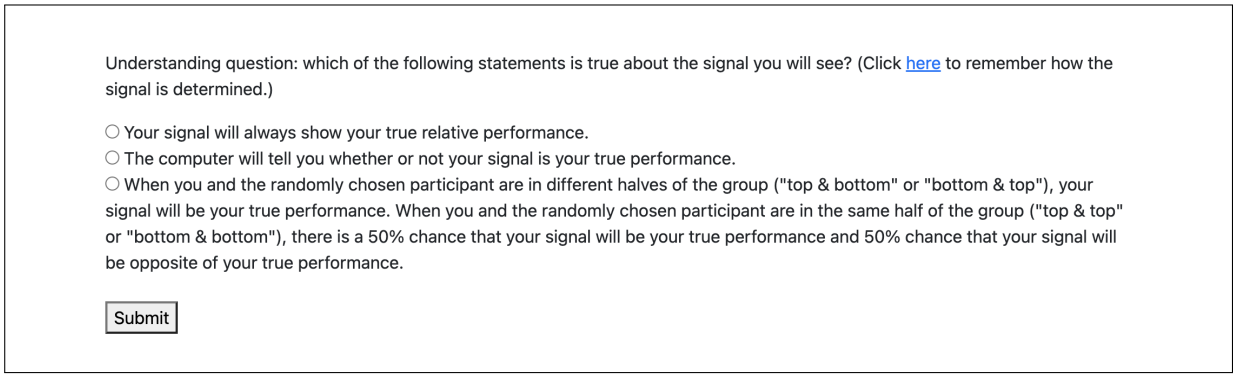


Figure A.13: Part II, Feedback if Bad News, Noisy Treatment

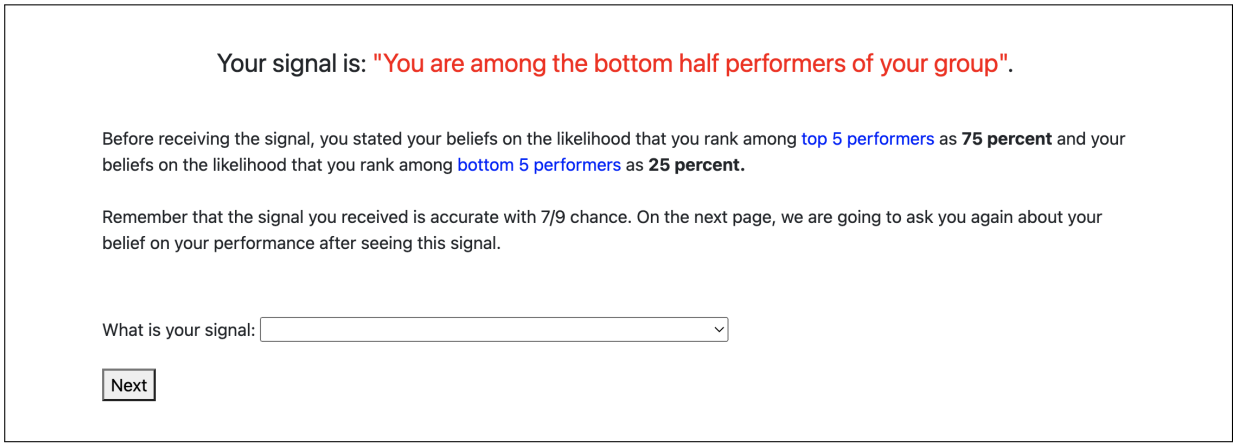


Figure A.14: Part II, Feedback if Bad News, Comparative Treatment

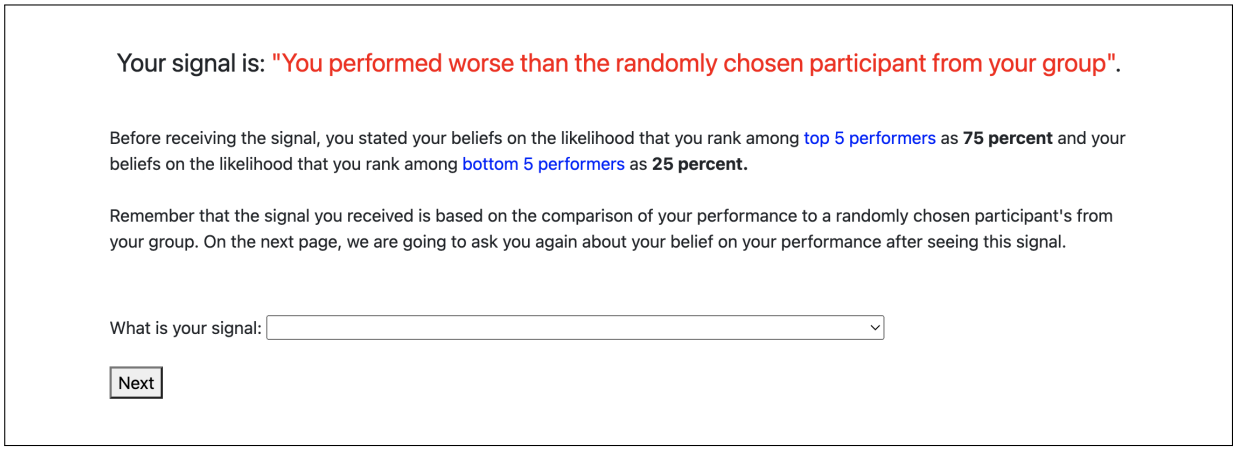


Figure A.15: Part II, Feedback if Bad News, NoisyComparative Treatment

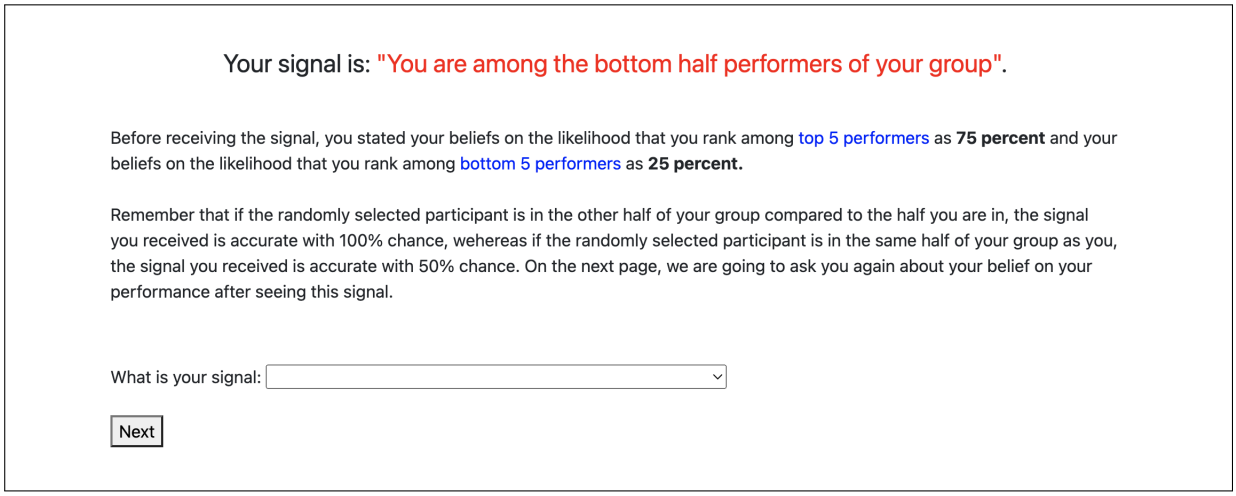


Figure A.16: Part II, Feedback if Good News, Noisy Treatment

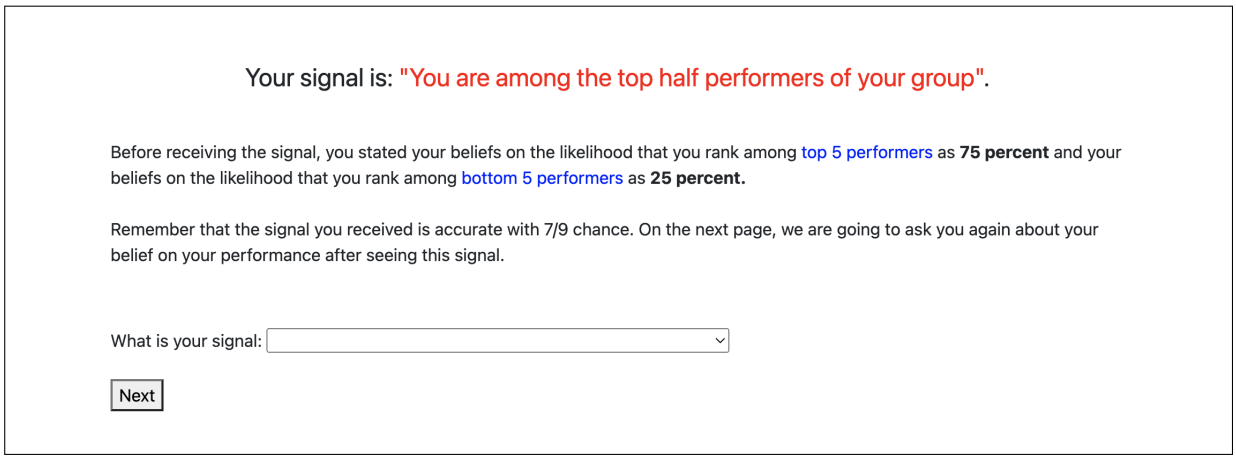


Figure A.17: Part II, Feedback if Good News, Comparative Treatment

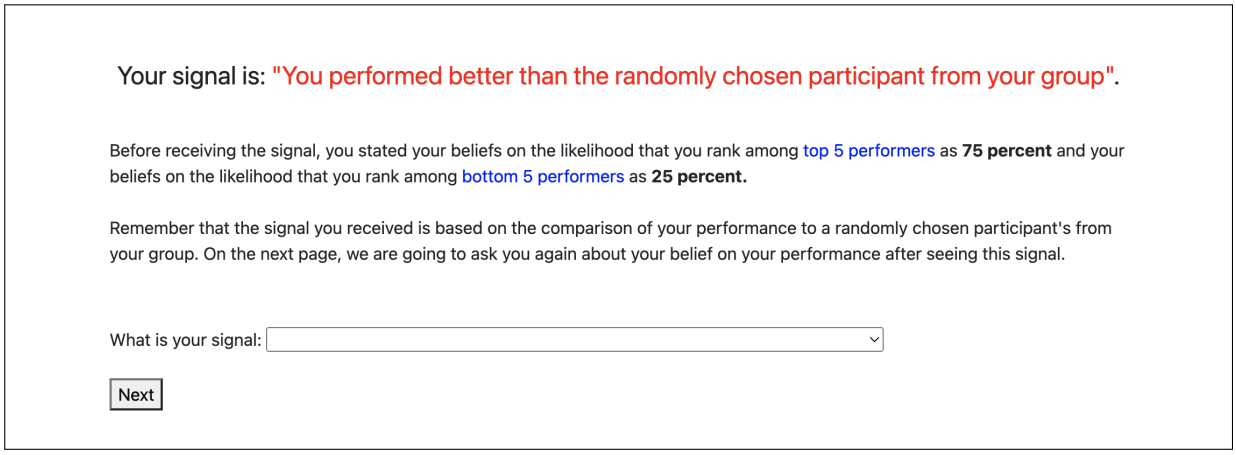


Figure A.18: Part II, Feedback if Good News, NoisyComparative Treatment

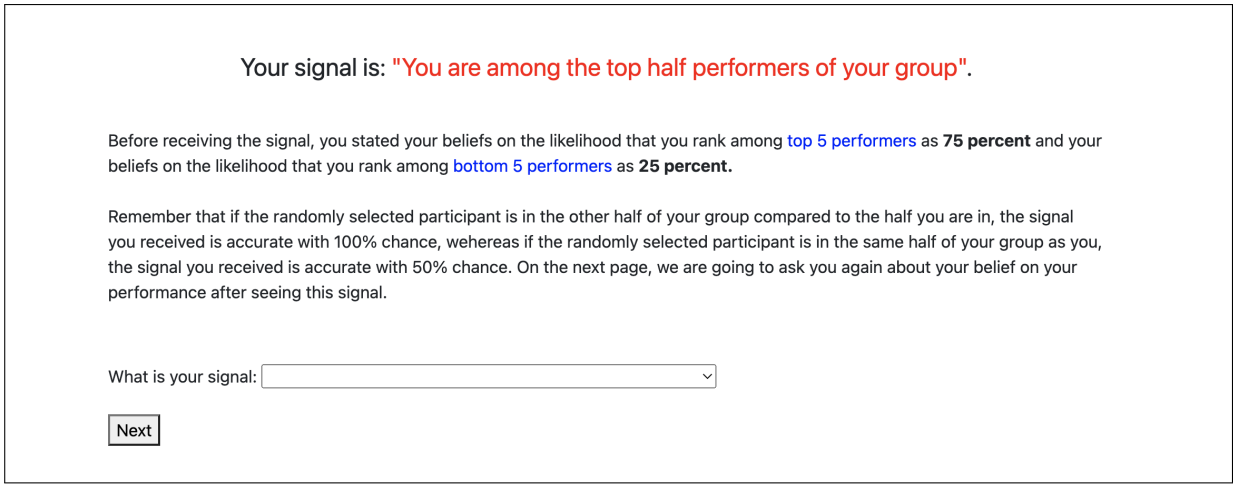


Figure A.19: Part II, Posterior Beliefs

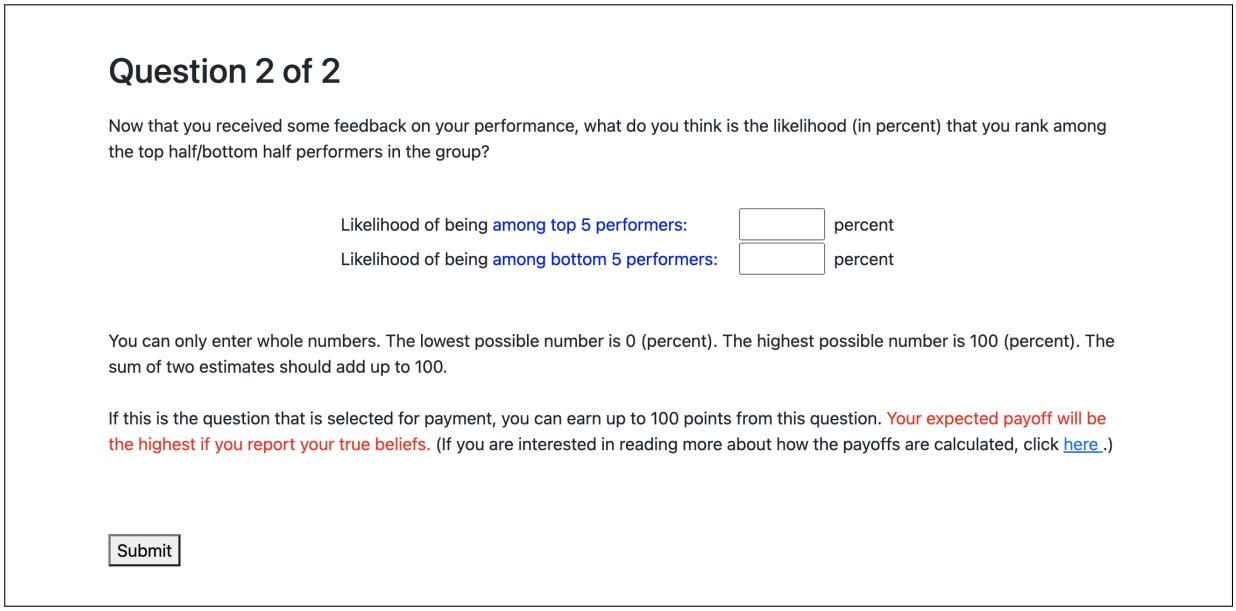
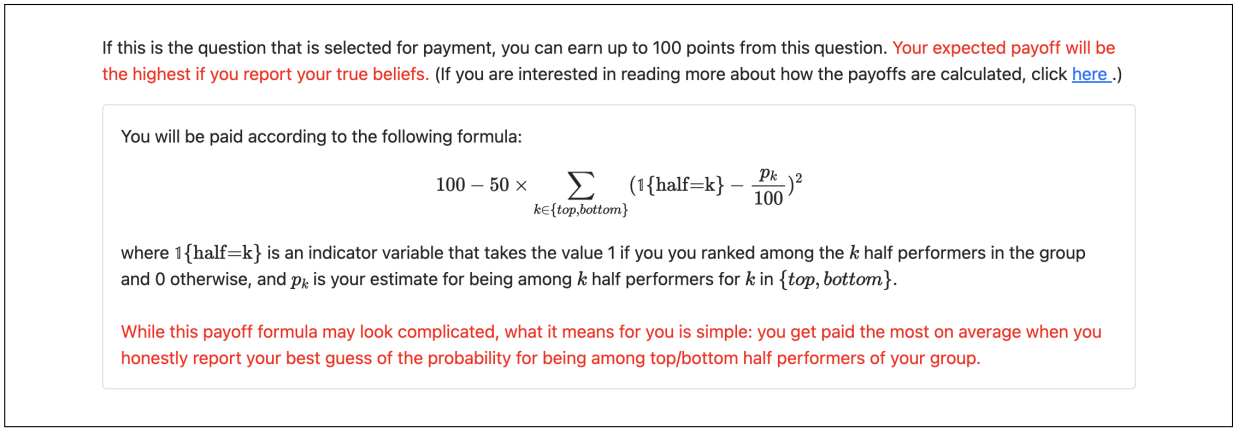


Figure A.20: Part II, Posterior Beliefs Pop-Up Box



Appendix B: Appendix to Chapter 2

B.1 Additional Tables and Figures

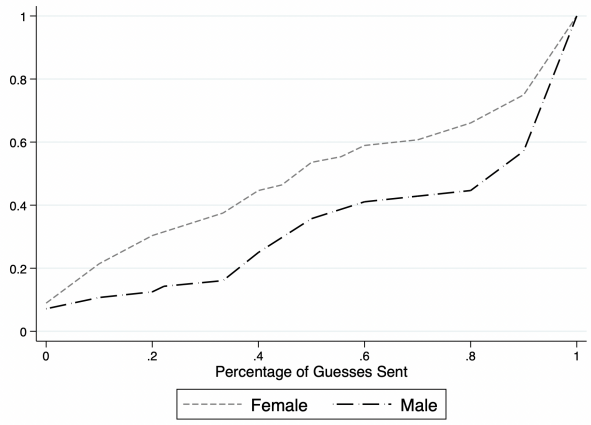
Table B.1: Order of Questions

Question	Number of red balls	Difficulty
1	18	Medium
2	6	Easy
3	59	Difficult
4	87	Medium
5	39	Difficult
6	81	Medium
7	40	Difficult
8	94	Easy
9	17	Medium
10	69	Difficult
11	24	Medium
12	62	Difficult
13	29	Medium
14	56	Difficult
15	95	Easy
16	49	Difficult
17	27	Medium
18	54	Difficult
19	85	Medium
20	9	Easy
21	76	Medium
22	34	Difficult
23	58	Difficult
24	7	Easy
25	79	Medium

Table B.2: Number of Subjects Broken Down by Gender, Treatment, and Role

	Senders		Receivers	
	Advice Treatment	Dictator Treatment	Advice Treatment	Dictator Treatment
Female	56	56	55	55
Male	56	56	56	56
Non-binary	0	1	1	2
N	112	113	112	113

Figure B.1: Cumulative Distribution Functions of Advice Sending by Gender for Difficult Questions



Notes: Figure plots the CDF of the percentage of difficult questions for which the senders sent their guess to the receiver in the Advice Treatment.

Table B.3: Sender Demographics Across Genders

	Male	Female	Difference	p-value
High Education	0.61	0.60	0.01	0.892
Employed	0.79	0.74	0.04	0.433
Age	37.0	39.8	-2.7*	0.068
Risk Averse	0.76	0.88	-0.12**	0.015
Rank Guess	1.76	2.04	-0.28***	0.008
N	112	112		

Notes: Table reports fraction of senders with Bachelor's degree or higher, fraction of senders who are employed, average sender age, fraction of senders who allocated less than their endowment to the risky project, and average self-perceived rank of guess accuracy among a random group of 4 senders. * p<0.1, ** p<0.05, *** p<0.01.

Table B.4: Probit Regressions Relating Sender's Guess-Sending to Gender for Difficult Questions, Advice Treatment

Guess Sent	(1)	(2)
Male	0.44**	0.42**
	(0.016)	(0.032)
Error		-0.02***
		(0.000)
Period		-0.01
		(0.135)
Risk Averse		-0.10
		(0.721)
High Education		0.13
		(0.511)
Employed		0.20
		(0.384)
Age		0.006
		(0.467)
Rank Guess: 2		-0.51**
		(0.027)
Rank Guess: 3		-0.86***
		(0.004)
Rank Guess: 4		0.43
		(0.510)
Constant	0.11	0.45
	(0.384)	(0.370)
N	1,106	1,106

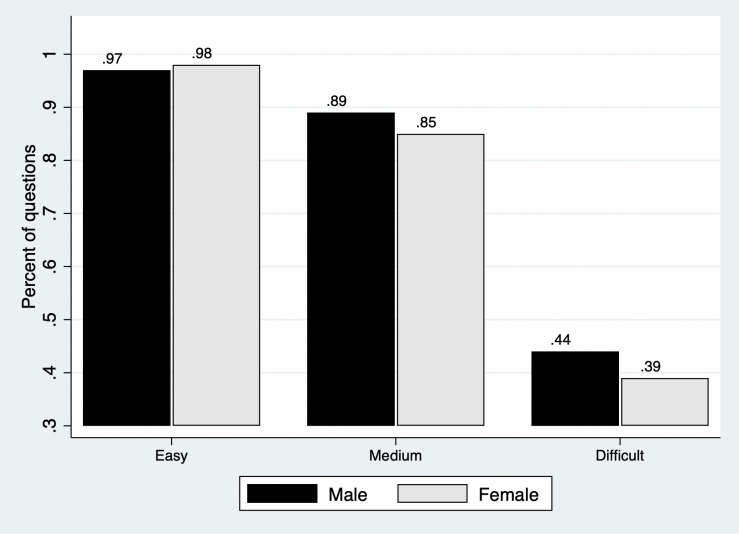
Notes: Dependent variable is *Guess Sent* (dummy variable equal to 1 if the sender sent their guess to the receiver in a given round and 0 otherwise). Control variables are *Male* (dummy variable equal to 1 for men and 0 for women), *Error* (normalized error of the sender in a given round), *Period* (round number), *Risk Averse* (dummy variable equal to 1 if subject allocated less than their endowment to the risky project task and 0 otherwise), *High Education* (dummy variable equal to 1 if subject's education is Bachelor's degree or higher and 0 otherwise), *Employed* (dummy variable equal to 1 if subject is employed and 0 otherwise), *Age*, and *Rank Guess* (indicator variables for subjects' self-confidence, takes values between 1-4). Errors are clustered at the individual level. p-values are reported in parentheses; * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Table B.5: Probit Regressions Relating Sender's Guess-Sending to Gender for Difficult Questions, Dictator Treatment

Guess Sent	(1)	(2)
Male	0.12	0.03
	(0.410)	(0.820)
Error		-0.02***
		(0.000)
Period		0.001
		(0.915)
Risk Averse		0.16
		(0.388)
High Education		-0.18
		(0.227)
Employed		-0.06
		(0.702)
Age		-0.004
		(0.570)
Rank Guess: 2		-0.82***
		(0.000)
Rank Guess: 3		-1.1***
		(0.000)
Rank Guess: 4		-0.56
		(0.404)
Constant	-0.29***	0.69*
	(0.005)	(0.063)
N	1,106	1,106

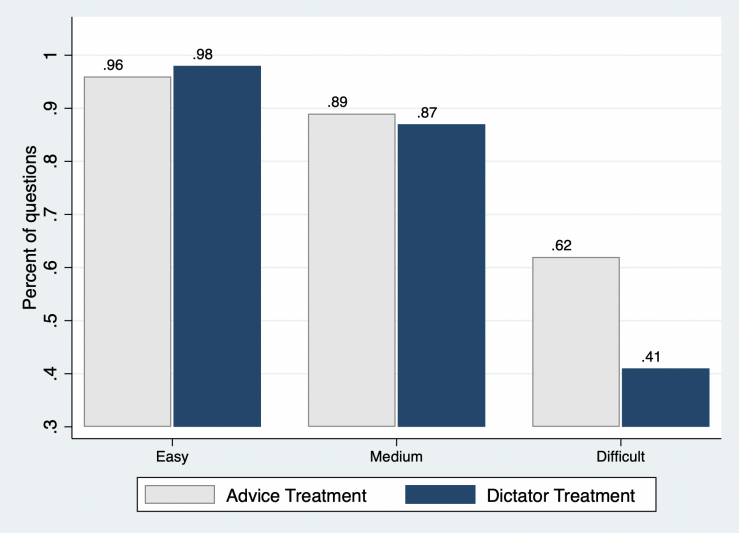
Notes: Dependent variable is *Guess Sent* (dummy variable equal to 1 if the sender sent their guess to the receiver in a given round and 0 otherwise). Control variables are *Male* (dummy variable equal to 1 for men and 0 for women), *Error* (normalized error of the sender in a given round), *Period* (round number), *Risk Averse* (dummy variable equal to 1 if subject allocated less than their endowment to the risky project task and 0 otherwise), *High Education* (dummy variable equal to 1 if subject's education is Bachelor's degree or higher and 0 otherwise), *Employed* (dummy variable equal to 1 if subject is employed and 0 otherwise), *Age*, and *Rank Guess* (indicator variables for subjects' self-confidence, takes values between 1-4). Errors are clustered at the individual level. p-values are reported in parentheses; * p<0.1, ** p<0.05, *** p<0.01.

Figure B.2: Percent of Questions For Which Senders Send Their Guess, Dictator Treatment



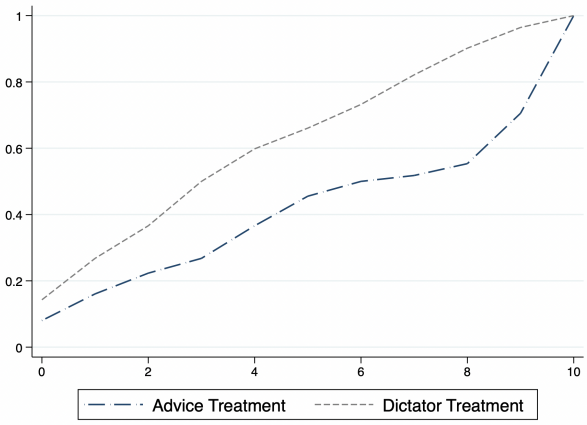
Notes: Figure illustrates percentages of easy, medium, and difficult questions for which the senders send their guess, broken down by gender. The p-values for the differences of percentages between men and women are $p = 0.565$ for easy, $p = 0.265$ for medium, and $p = 0.478$ for difficult questions.

Figure B.3: Percent of Questions For Which the Senders Send Their Guess Across Treatments



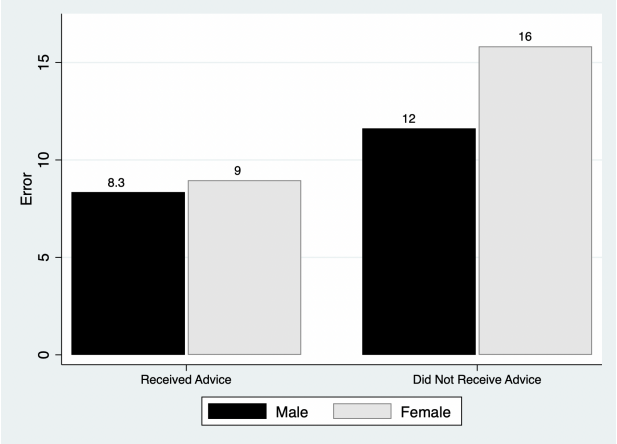
Notes: Figure illustrates percentages of easy, medium, and difficult questions for which the senders send their guess, broken down by treatment. The p-values for the differences of percentages between the Advice and Dictator Treatments are $p = 0.592$ for easy, $p = 0.053$ for medium, and $p < 0.001$ for difficult questions.

Figure B.4: Cumulative Distribution Functions of Guess Sending by Treatment for Difficult Questions



Notes: Figure plots the CDFs of percentages of difficult questions for which the senders send their guess to the receiver.

Figure B.5: Performance of Receivers for Difficult Questions, Advice Treatment, Broken Down By Gender and Presence of Advice



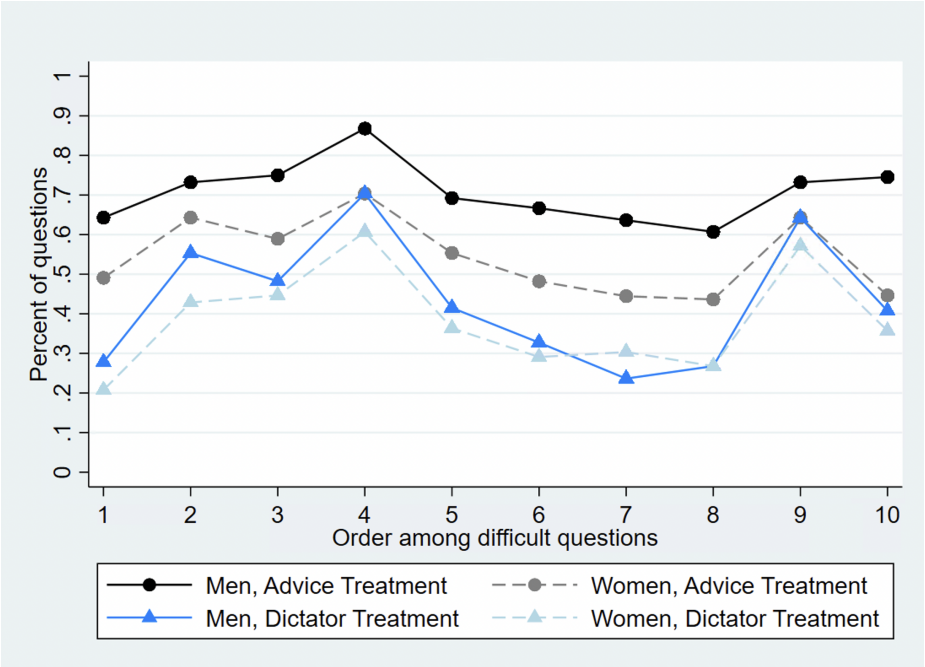
Notes: Figure illustrates average normalized errors of receivers, broken down by gender and presence of advice for difficult questions in the Advice Treatment. The p-values for the difference in performances between men and women are $p = 0.4500$ for the rounds in which the receivers received advice and $p = 0.0097$ for the rounds in which the receivers did not receive advice.

Table B.6: OLS Regressions Relating Receivers' Performance to Senders' Guess Sending for Difficult Questions

	Advice Treatment		Dictator Treatment	
Error	(1)	(2)	(3)	(4)
Guess Sent	-4.5***	-4.4***	-0.64	-0.52
	(0.000)	(0.000)	(0.328)	(0.410)
Male		-0.9		-1.4
		(0.401)		(0.123)
Rank Guess = 2		4.1***		1.1
		(0.001)		(0.427)
Rank Guess = 3		5.1***		1.4
		(0.002)		(0.386)
Rank Guess = 4		1.7		7.8**
		(0.615)		(0.015)
Period		-0.02		-0.05
		(0.756)		(0.424)
Risk Averse		0.43		1.0
		(0.668)		(0.331)
High Education		-0.08		0.003
		(0.953)		(0.998)
Employed		2.2**		1.9*
		(0.045)		(0.064)
Age		0.01		0.03
		(0.800)		(0.243)
Constant	13***	7.7***	11***	7.6***
	(0.000)	(0.005)	(0.000)	(0.002)
Controls	No	Yes	No	Yes
N	1,104	1,104	1,105	1,105

Notes: Dependent variable is *Error* (normalized error of the receiver in a given round). Control variables are *Guess Sent* (dummy variable equal to 1 if the sender sent their guess in a given round), *Male* (dummy variable equal to 1 for men and 0 for women), *Rank Guess* (indicator variables for subjects' self-confidence, takes values between 1-4), *Period* (round number), *Risk Averse* (dummy variable equal to 1 if subject allocated less than their endowment to the risky project task and 0 otherwise), *High Education* (dummy variable equal to 1 if subject's education is Bachelor's degree or higher and 0 otherwise), *Employed* (dummy variable equal to 1 if subject is employed and 0 otherwise), and *Age*. Errors are clustered at the individual level. p-values are reported in parentheses; * p<0.1, ** p<0.05, *** p<0.01.

Figure B.6: Percentage of Guesses Sent in Difficult Questions, Broken Down By Gender, Treatment, and Order of the Question among other Difficult Questions



Notes: Figure illustrates percentage of questions for which senders sent their guess in difficult questions, broken down by gender, treatment, and the order of the question among difficult questions. (The order of each difficult question ordered 1-10 in this graph correspond to questions (3,5,7,10,12,14,16,18,22,23) among all questions).

Table B.7: Percentage of Advice Sending in Difficult Questions by Gender for Different Values of $\overline{\delta}$

$\overline{\delta}$	N_D	Men	Women	Diff	p-value	$\overline{\delta}$	N_D	Men	Women	Diff	p-value
0	25	0.826	0.766	0.060	0.068	25	12	0.738	0.593	0.145	0.017
1	25	0.826	0.766	0.060	0.068	26	12	0.738	0.593	0.145	0.017
2	25	0.826	0.766	0.060	0.068	27	11	0.725	0.569	0.156	0.014
3	25	0.826	0.766	0.060	0.068	28	11	0.725	0.569	0.156	0.014
4	25	0.826	0.766	0.060	0.068	29	10	0.706	0.542	0.165	0.014
5	24	0.822	0.757	0.065	0.058	30	10	0.706	0.542	0.165	0.014
6	22	0.810	0.736	0.074	0.052	31	9	0.688	0.525	0.163	0.022
7	21	0.806	0.723	0.082	0.047	32	9	0.688	0.525	0.163	0.022
8	21	0.806	0.723	0.082	0.047	33	9	0.688	0.525	0.163	0.022
9	20	0.800	0.710	0.090	0.041	34	8	0.682	0.510	0.172	0.018
10	20	0.800	0.710	0.090	0.041	35	8	0.682	0.510	0.172	0.018
11	20	0.800	0.710	0.090	0.041	36	8	0.682	0.510	0.172	0.018
12	20	0.800	0.710	0.090	0.041	37	8	0.682	0.510	0.172	0.018
13	19	0.795	0.703	0.093	0.043	38	7	0.682	0.504	0.178	0.014
14	19	0.795	0.703	0.093	0.043	39	6	0.674	0.480	0.193	0.012
15	18	0.789	0.691	0.098	0.043	40	5	0.658	0.459	0.199	0.011
16	18	0.789	0.691	0.098	0.043	41	4	0.662	0.452	0.210	0.008
17	17	0.780	0.675	0.105	0.039	42	3	0.634	0.455	0.179	0.027
18	16	0.776	0.660	0.115	0.032	43	3	0.634	0.455	0.179	0.027
19	15	0.769	0.644	0.126	0.027	44	2	0.625	0.446	0.179	0.047
20	15	0.769	0.644	0.126	0.027	45	2	0.625	0.446	0.179	0.047
21	14	0.759	0.627	0.132	0.025	46	1	0.636	0.455	0.182	0.084
22	14	0.759	0.627	0.132	0.025	47	1	0.636	0.455	0.182	0.084
23	14	0.759	0.627	0.132	0.025	48	1	0.636	0.455	0.182	0.084
24	12	0.738	0.593	0.145	0.017						

Notes: $\overline{\delta}$ is the cutoff for determining question difficulty such that questions with $\delta > \overline{\delta}$ are classified as "difficult". N_D is the number of questions classified as difficult when $\overline{\delta}$ is the cutoff for question difficulty. Columns *Men* and *Women* depict the percentage of questions for which men and women senders sent their guess in difficult questions in the Advice Treatment when difficulty of questions are determined by the corresponding $\overline{\delta}$ cutoff. *Diff* is the difference in the percent advice sent by men and women. *p-value* is calculated using a Mann Whitney U-test.

Table B.8: Probit Regressions Relating Sender's Guess-Sending to Gender and Difficulty Index, Advice Treatment

Guess Sent	(1)	(2)
Male	-0.531*	-0.549*
	(0.055)	(0.062)
Delta	-0.054***	-0.052***
	(0.000)	(0.000)
Male×Delta	0.024***	0.024***
	(0.001)	(0.002)
Period		0.000
		(0.959)
Error		-0.012***
		(0.000)
Risk Averse		-0.168
		(0.466)
High Education		0.119
		(0.479)
Employed		0.252
		(0.222)
Age		0.004
		(0.612)
Rank Guess=2		-0.408**
		(0.042)
Rank Guess=3		-0.762***
		(0.002)
Rank Guess=4		0.361
		(0.566)
Constant	2.300***	2.494***
	(0.000)	(0.000)
N	2,773	2,773

Notes: Dependent variable is *Guess Sent* (dummy variable equal to 1 if the sender sent their guess to the receiver in a given round and 0 otherwise). Control variables are *Male* (dummy variable equal to 1 for men and 0 for women), *Delta* (difficulty index, $\delta = (100 - |b - r|)/2$, where b and r correspond to the number of red and blue balls in the box, respectively), *Male*×*Delta* interaction term, *Period* (round number), *Error* (normalized error of the sender in a given round), *Risk Averse* (dummy variable equal to 1 if subject allocated less than their endowment to the risky project task and 0 otherwise), *High Education* (dummy variable equal to 1 if subject's education is Bachelor's degree or higher and 0 otherwise), *Employed* (dummy variable equal to 1 if subject is employed and 0 otherwise), *Age*, and *Rank Guess* (indicator variables for subjects' self-confidence, takes values between 1-4). Errors are clustered at the individual level. p-values are reported in parentheses; * p<0.1, ** p<0.05, *** p<0.01.

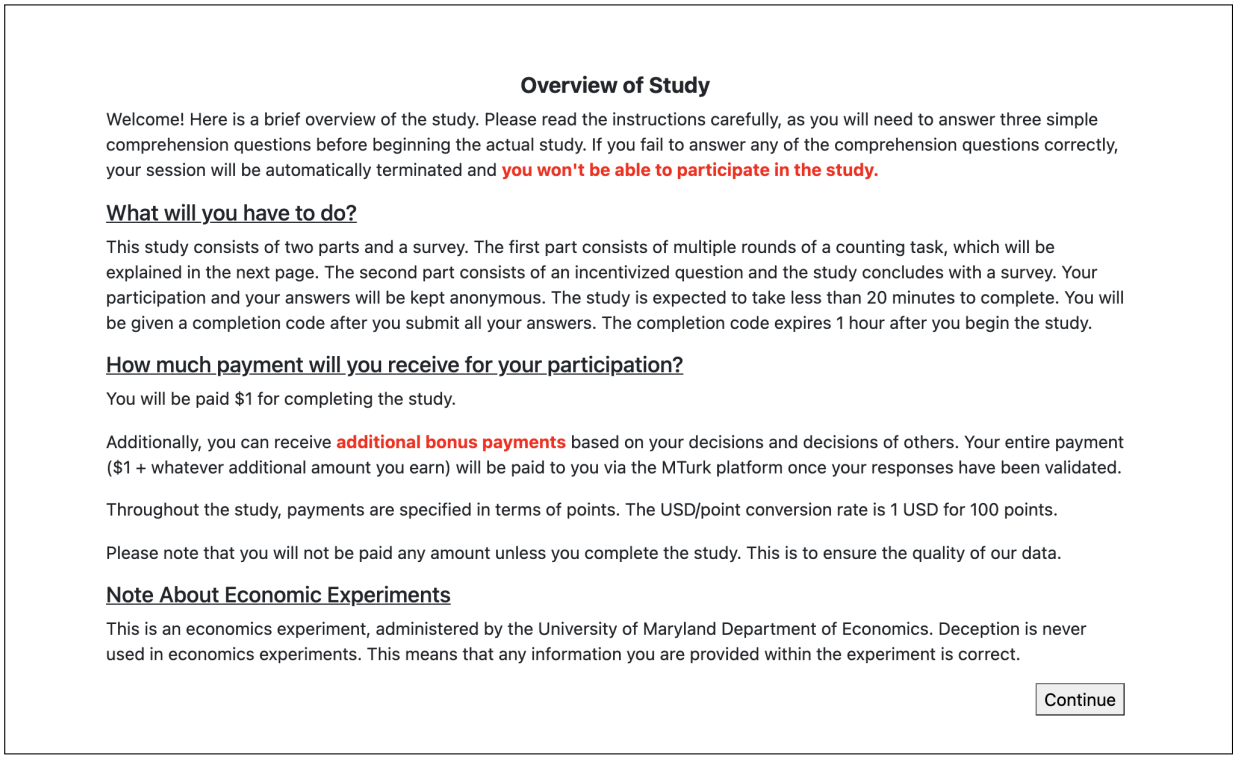
Table B.9: Probit Regressions Relating Sender's Guess-Sending to Gender and Difficulty Level, Advice Treatment

Guess Sent	(1)	(2)
Male	-0.730**	-0.790**
	(0.021)	(0.013)
Medium	-1.027***	-1.026***
	(0.000)	(0.000)
Difficult	-2.081***	-2.061***
	(0.000)	(0.000)
Male× Medium	0.814***	0.865***
	(0.002)	(0.001)
Male× Difficult	1.167***	1.195***
	(0.001)	(0.000)
Period		-0.007*
		(0.057)
Error		-0.013***
		(0.000)
Risk Averse		-0.170
		(0.462)
High Education		0.126
		(0.452)
Employed		0.252
		(0.220)
Age		0.004
		(0.597)
Rank Guess=2		-0.405**
		(0.043)
Rank Guess=3		-0.762***
		(0.002)
Rank Guess=4		0.343
		(0.582)
Constant	2.189***	2.491***
	(0.000)	(0.000)
Gender difference in advice sending for each difficulty level:		
Easy	-0.730**	-0.790**
	(0.021)	(0.013)
Medium	0.084	0.075
	(0.649)	(0.708)
Difficult	0.437**	0.405**
	(0.016)	(0.036)
N	2,773	2,773

Notes: The first half of the table reports coefficients from regressing *Guess Sent* (dummy variable equal to 1 if the sender sent their guess to the receiver in a given round and 0 otherwise) on each relevant variable with p-values in parentheses. Errors are clustered at the individual level. p-values are reported in parentheses; * p¡0.1, ** p¡0.05, *** p¡0.01. The second half of the table reports gender difference in advice sending for each difficulty level. *Easy* reports the difference $(\beta_{Constant} + \beta_{Male}) - \beta_{Constant}$ and the p-value from Chow-test on equality of this expression to 0. Similarly, *Medium* reports the difference $(\beta_{Constant} + \beta_{Male} + \beta_{Medium} + \beta_{Male \times Medium}) - (\beta_{Constant} + \beta_{Medium})$, and *Difficult* reports the difference $(\beta_{Constant} + \beta_{Male} + \beta_{Difficult} + \beta_{Male \times Difficult}) - (\beta_{Constant} + \beta_{Difficult})$ and the relevant p-values.

B.2 Instructions

Figure B.7: Welcome Page



Overview of Study

Welcome! Here is a brief overview of the study. Please read the instructions carefully, as you will need to answer three simple comprehension questions before beginning the actual study. If you fail to answer any of the comprehension questions correctly, your session will be automatically terminated and **you won't be able to participate in the study.**

What will you have to do?

This study consists of two parts and a survey. The first part consists of multiple rounds of a counting task, which will be explained in the next page. The second part consists of an incentivized question and the study concludes with a survey. Your participation and your answers will be kept anonymous. The study is expected to take less than 20 minutes to complete. You will be given a completion code after you submit all your answers. The completion code expires 1 hour after you begin the study.

How much payment will you receive for your participation?

You will be paid $1 for completing the study.

Additionally, you can receive **additional bonus payments** based on your decisions and decisions of others. Your entire payment ($1 + whatever additional amount you earn) will be paid to you via the MTurk platform once your responses have been validated.

Throughout the study, payments are specified in terms of points. The USD/point conversion rate is 1 USD for 100 points.

Please note that you will not be paid any amount unless you complete the study. This is to ensure the quality of our data.

Note About Economic Experiments

This is an economics experiment, administered by the University of Maryland Department of Economics. Deception is never used in economics experiments. This means that any information you are provided within the experiment is correct.

Continue

Figure B.8: Advice Treatment, Senders

Instructions for the Counting Task

In this part of the experiment, you will participate in 25 independent decision periods, in each of which you will see a box containing 100 red and blue balls. Your role in the experiment is **advisor.** Once you complete the experiment, your answers will be matched to participants who are assigned the role of **decision maker.** The decision maker you will be matched with will be determined randomly for each round, so you will be **matched with a new decision maker in each round.**

Both you and the decision maker you are matched within a given round will see the same box, which will disappear after 10 seconds. The task is to submit the best estimate about the number of red balls in the box.

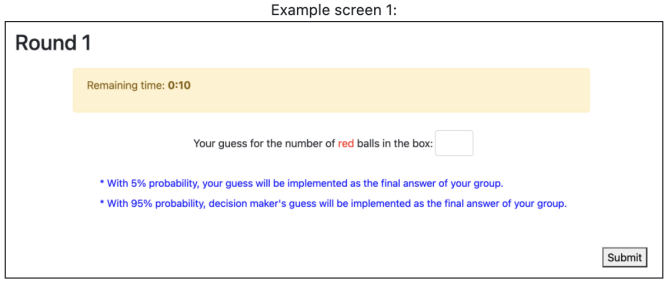
What will you do?

You will enter your guess about the number of red balls in the box. Then, you will choose whether to send your guess as an advice to the decision maker. If you choose to send advice, the decision maker will see your answer before submitting theirs. You may also choose not to send advice.

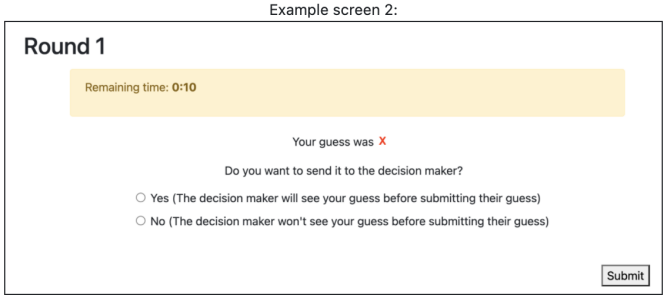
In each round, there is a 5% chance that your guess will be implemented as the final answer of your group. With 95% probability, the decision maker's guess will be implemented as the final answer of your group.

You will have 10 seconds to make each decision. **You will not be eligible for payment** if you fail to submit your response to three or more questions within the allocated time. This is to ensure that you are paying attention.

Example screen 1:



Example screen 2:



What will the decision maker do?

The decision maker will enter their guess about the number of red balls in the same box. The decision maker will also have 10 seconds to submit an answer. If you choose to send advice, the decision maker will see your advice before submitting their guess, otherwise they will be informed that you didn't send advice. It is up to the decision maker whether they consider your advice or not.

How are payoffs calculated?

Your payoff in each round is equal to the decision maker's payoff. In each round, there is a 5% probability that your guess is implemented as the group's final answer. Otherwise, the final answer to determine payoff of your group is the decision maker's answer. The payoff is higher the closer the final answer is to the correct number of red balls in the box. The exact payoffs are reported below ("Error" is the difference between the final answer and the correct number of red balls, 100 points = 1 USD):

Error	Payoff
0	400 points
1-3	200 points
4-10	100 points
11-15	50 points
greater than 15	0 points

What will be my earnings?

At the end of the experiment, one of the 25 rounds will be randomly selected for payment. Your earning for this part of the experiment will be equal to your payoff in the randomly selected round.

Continue

Figure B.9: Advice Treatment, Receivers

Instructions for the Counting Task

In this part of the experiment, you will participate in 25 independent decision periods, in each of which you will see a box containing 100 red and blue balls. Your role in the experiment is **decision maker.** During the experiment, you will be matched to participants who previously completed this study and were assigned the role of **advisor.** The advisor you will be matched with will be determined randomly for each round, so you will be **matched with a new advisor in each round.**

In each round, you will see a different box that will disappear in 10 seconds. The advisor you are matched within a given round also saw the same box for 10 seconds. The task is to submit the best estimate about the number of red balls in the box.

What did the advisor do?

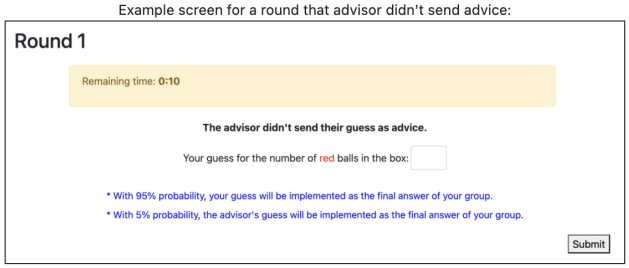
The advisor entered their guess about the number of red balls in the box and then chose whether to send it as an advice to you. If they chose to send advice, you will see their answer before submitting yours. They may have also chosen not to send advice. The advisor had 10 seconds to make each decision. In each round, there is a 5% chance that the advisor's guess is implemented as the final answer of your group. With 95% probability, your guess will be implemented as the final answer of your group.

What will you do?

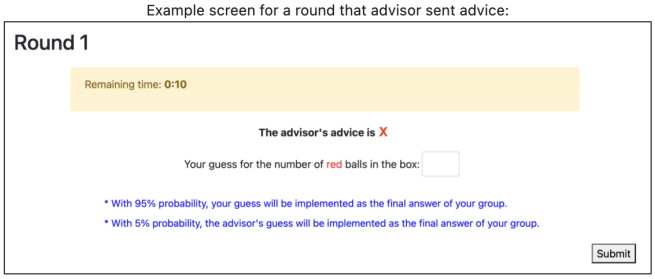
You will enter your guess about the number of red balls in the same box. You also have 10 seconds to submit an answer. **You will not be eligible for payment** if you fail to submit your response to three or more questions within the allocated time. This is to ensure that you are paying attention.

If the advisor chose to send advice, you will see their advice before submitting your answer, otherwise you will be informed that they didn't send advice. It is up to you whether you consider their advice or not.

Example screen for a round that advisor didn't send advice:



Example screen for a round that advisor sent advice:



How are payoffs calculated?

The advisor's payoff in each round will be equal to your payoff. (The advisors will receive payment after the decision makers complete the study). In each round, there is a 5% probability that the advisor's guess is implemented as the group's final answer. With 95% probability, the final answer to determine payoff of your group is your guess. The payoff is higher the closer the final answer is to the correct number of red balls in the box. The exact payoffs are reported below ("Error" is the difference between the final answer and the correct number of red balls, 100 points = 1 USD):

Error	Payoff
0	400 points
1-3	200 points
4-10	100 points
11-15	50 points
greater than 15	0 points

What will be my earnings?

At the end of the experiment, one of the 25 rounds will be randomly selected for payment. Your earning for this part of the experiment will be equal to your payoff in the randomly selected round.

Continue

Figure B.10: Dictator Treatment, Senders

Instructions for the Counting Task

In this part of the experiment, you will participate in 25 independent decision periods, in each of which you will see a box containing 100 red and blue balls. Your role in the experiment is **advisor.** Once you complete the experiment, your answers will be matched to participants who are assigned the role of **decision maker.** The decision maker you will be matched with will be determined randomly for each round, so you will be **matched with a new decision maker in each round.**

Both you and the decision maker you are matched within a given round will see the same box, which will disappear after 10 seconds. The task is to submit the best estimate about the number of red balls in the box.

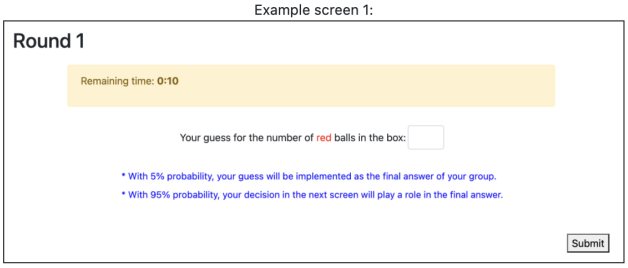
What will you do?

You will enter your guess about the number of red balls in the box. Then, you will choose whether to send your guess to be implemented as the final answer of your group or let the decision maker's guess be implemented as the final answer of your group.

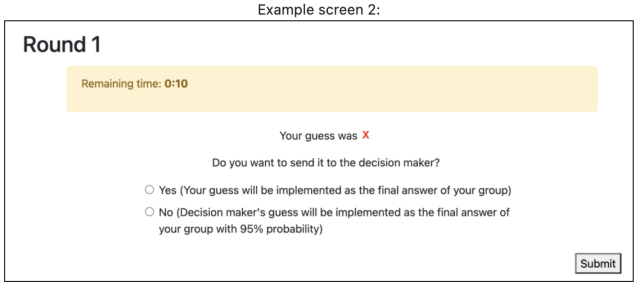
In each round, there is a 5% chance that your guess will be implemented as the final answer of your group regardless of your choice. With 95% probability, your choice will determine whose guess will be implemented as the final answer.

You will have 10 seconds to make each decision. **You will not be eligible for payment** if you fail to submit your response to three or more questions within the allocated time. This is to ensure that you are paying attention.

Example screen 1:



Example screen 2:



What will the decision maker do?

The decision maker will enter their guess about the number of red balls in the same box. They will also have 10 seconds to submit an answer. Once the decision maker submits their answer, they will be informed of your choice to send your guess. If you choose to send your guess, the decision maker will see your guess and will be informed that your decision will be implemented as the final answer of your group. If you didn't send your guess, they will be informed that with 95% probability their guess will be implemented as the final answer of your group and with 5% probability your guess will be implemented as the final answer of your group.

How are payoffs calculated?

Your payoff in each round is equal to the decision maker's payoff. In each round, there is a 5% probability that your guess is implemented as the group's final answer regardless of your choice. Otherwise, the final answer to determine payoff of your group is your guess if you chose to send your guess and it is the decision maker's guess if you chose not to send your guess. The payoff is higher the closer the final answer is to the correct number of red balls in the box. The exact payoffs are reported below ("Error" is the difference between the final answer and the correct number of red balls, 100 points = 1 USD):

Error	Payoff
0	400 points
1-3	200 points
4-10	100 points
11-15	50 points
greater than 15	0 points

What will be my earnings?

At the end of the experiment, one of the 25 rounds will be randomly selected for payment. Your earning for this part of the experiment will be equal to your payoff in the randomly selected round.

Continue

Figure B.11: Dictator Treatment, Receivers

Instructions for the Counting Task

In this part of the experiment, you will participate in 25 independent decision periods, in each of which you will see a box containing 100 red and blue balls. Your role in the experiment is **decision maker.** During the experiment, you will be matched to participants who previously completed this study and were assigned the role of **advisor.** The advisor you will be matched with will be determined randomly for each round, so you will be **matched with a new advisor in each round.**

In each round, you will see a different box that will disappear in 10 seconds. The advisor you are matched within a given round also saw the same box for 10 seconds. The task is to submit the best estimate about the number of red balls in the box.

What did the advisor do?

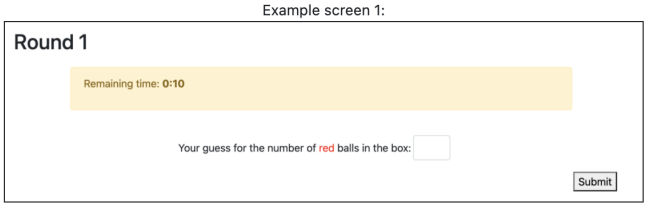
The advisor entered their guess about the number of red balls in the box. Then, they chose whether to send their guess to be implemented as the final answer of your group or let you submit your guess as the final answer of your group. In each round, there is a 5% chance that the advisor's guess is implemented as the final answer of your group regardless of their choice. With 95% probability, the advisor's choice will determine whose guess will be implemented as the final answer.

What will you do?

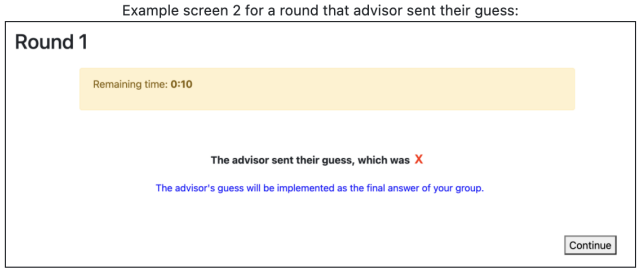
You will enter your guess about the number of red balls in the same box. You also have 10 seconds to submit an answer. **You will not be eligible for payment** if you fail to submit your response to three or more questions within the allocated time. This is to ensure that you are paying attention.

With 95% probability, the advisor's choice determines whose answer is implemented as the final decision of your group: the advisor's answer is implemented as the final answer of your group if they chose to send their guess, your answer is implemented as the final answer of your group if they didn't send their guess. With 5% probability, the advisor's guess is implemented as the final answer of your group regardless of their choice.

Example screen 1:



Example screen 2 for a round that advisor sent their guess:



Example screen 2 for a round that advisor didn't send their guess:

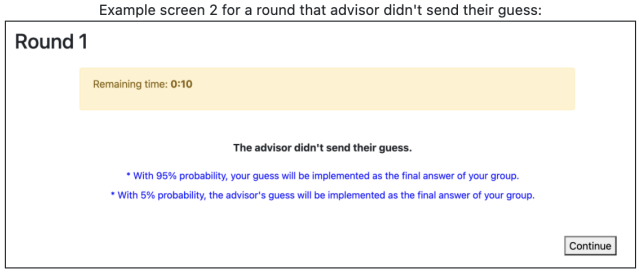


Figure B.12: Dictator Treatment, Receivers (Continued)

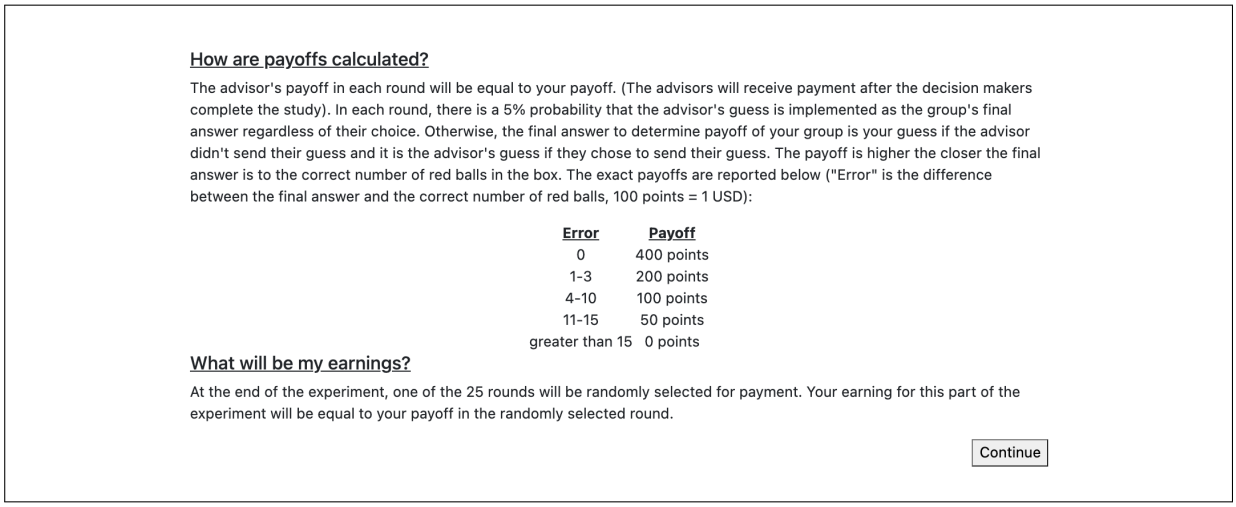
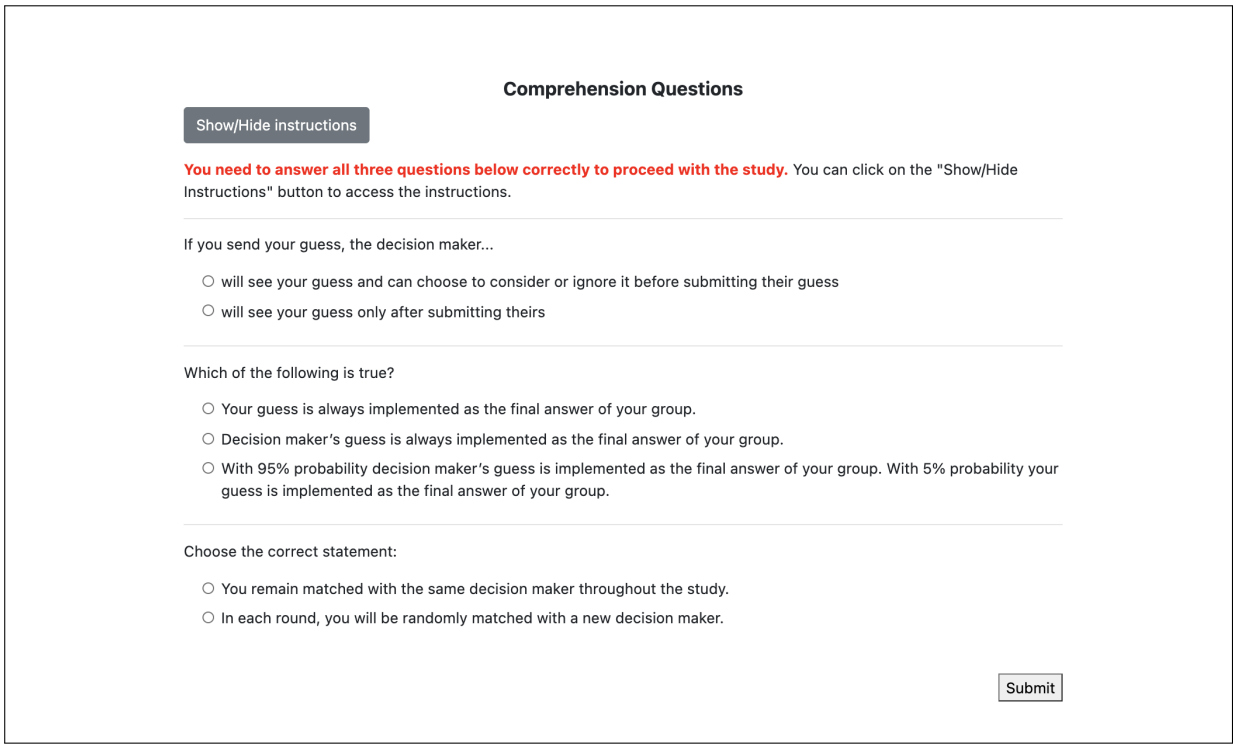
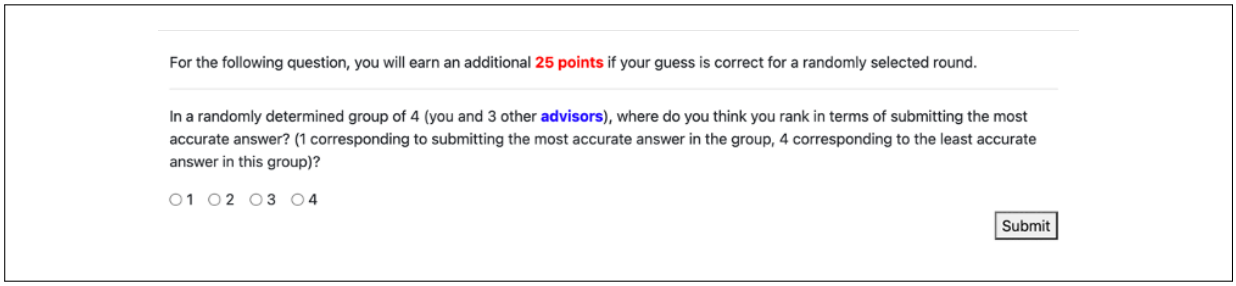


Figure B.13: Comprehension Questions



Notes: Above screenshot is the comprehension questions for senders in the Advice Treatment. The comprehension questions in the Dictator Treatment for the senders is the same except the third answer of the second question is "You choose whether your guess or the receiver's guess is implemented as the final answer of your group. With 95% probability the final answer is determined in accordance with your choice. With 5% probability your guess is implemented as the final answer of your group independent of whether you sent your guess or not". The only difference in comprehension questions for the receivers in both treatments is in wording to address for the correct role.

Figure B.14: Confidence Elicitation Question



Notes: Above screenshot is the confidence elicitation question for the senders. Receivers answered the same question except they were grouped with 3 other receivers, not senders.

Figure B.15: Risk Elicitation Question

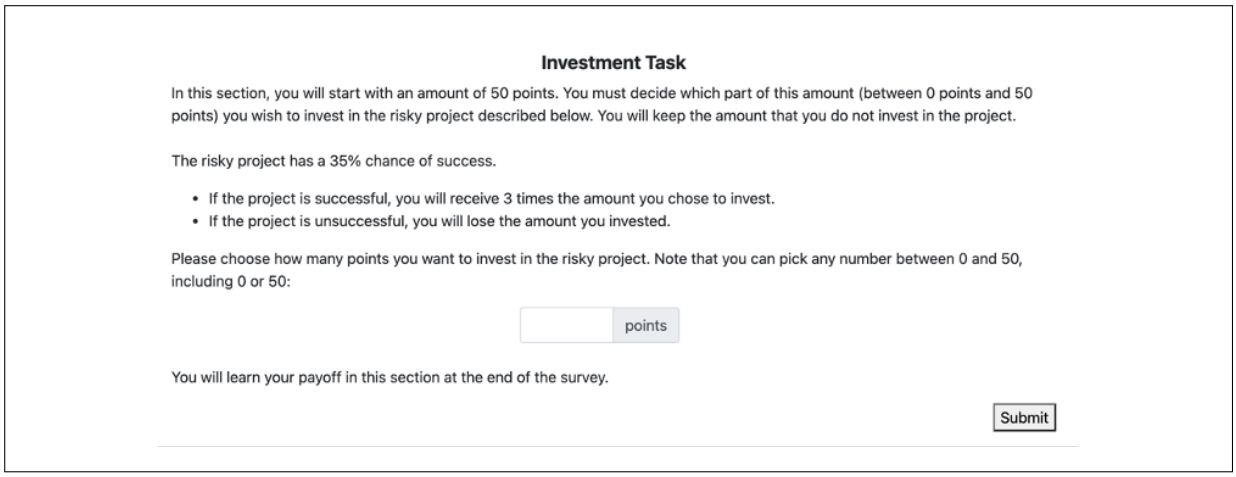
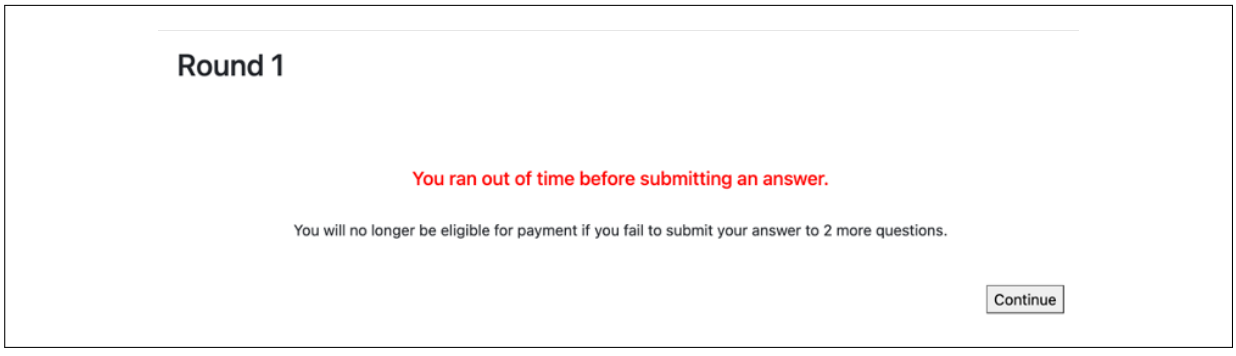


Figure B.16: Timeout Screen



Appendix C: Appendix to Chapter 3

C.1 Additional Tables

Table C.1: Types of an Agent

Type t	**Value** $v(t)$	**Probability** q_t	**Available Messages**
High	100	50%	{"I don't have evidence for my type"}
Low	0	50%	{"My type is low", "I don't have evidence for my type"}

Table C.2: Probit Regressions Relating Withholding Information to the Difference Between Rewards in the Commitment Treatment Conditioning on the Difference being Positive

	(1)	(2)	(3)
Difference Between	0.022***	0.023***	0.019**
Rewards	(0.009)	(0.009)	(0.011)
Reward for			-0.007*
Low Evidence			(0.083)
Period		0.026**	0.019*
		(0.014)	(0.081)
Gender		-0.263	-0.289
		(0.263)	(0.194)
Risk Aversion		-0.848	-0.871
		(0.159)	(0.173)
Ability to		0.114	0.067
Bayesian update		(0.784)	(0.865)
Constant	0.046	0.698	1.155
	(0.814)	(0.110)	(0.129)
Observations	320	320	320

Notes: Dependent variable *withhold evidence* is equal to 1 if the low type agent sent no evidence in the Commitment treatment and 0 if they sent low evidence. *Difference Between Rewards* is the difference between Reward for No Evidence and Reward for Low Evidence. *Period* takes values from 1 to 20 and represents the period. *Gender* is a dummy variable that takes the value 1 if subject is female and 0 otherwise. *Risk Aversion* takes the value 1 if the subject is classified as risk averse based on the number of safe options they chose in Activity 1 and 0 otherwise. *Ability to Bayesian update* is a dummy variable that takes the value 1 if subject answered the Activity 2 question of Part II correctly and 0 otherwise. p-values computed by score wild bootstrap procedure are in parentheses (clustered at the session level); * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

C.2 Regressions Without Bootstrapping Procedure

Table C.3: Tobit Regressions Relating Reward for No-Evidence to Treatment

	(1)	(2)
Commitment	15.32**	15.06**
	(0.026)	(0.029)
Period		-0.47
		(0.213)
Gender		-1.6
		(0.853)
Risk aversion		-0.89
		(0.897)
Ability to		-7.0
Bayesian update		(0.427)
Constant	50.3***	59.9***
	(0.000)	(0.000)
Observations	1,233	1,233

Notes: Dependent variable is *reward for no evidence*, bounded between 0 and 100. *Commitment* is a dummy variable that takes the value 1 if subject is in Commitment treatment and 0 if subject is in No-Commitment treatment. *Period* takes values from 1 to 20 and represents the period. *Gender* is a dummy variable that takes the value 1 if subject is female and 0 otherwise. *Risk Aversion* takes the value 1 if the subject is classified as risk averse based on the number of safe options they chose in Activity 1 and 0 otherwise. *Ability to Bayesian update* is a dummy variable that takes the value 1 if subject answered the Activity 2 question of Part II correctly and 0 otherwise. Standard errors are clustered at the individual level. p-values are in parentheses; * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Table C.4: Probit Regressions Relating Withholding Information to the Rewards in the Commitment Treatment Conditioning on the Difference being Positive

	(1)	(2)
Reward for	0.018***	0.019***
No Evidence	(0.000)	(0.000)
Reward for	-0.026***	-0.027***
Low Evidence	(0.000)	(0.000)
Period		0.019
		(0.204)
Gender		-0.289
		(0.275)
Risk aversion		-0.871***
		(0.000)
Ability to		0.067
Bayesian update		(0.799)
Constant	0.382**	1.155***
	(0.031)	(0.008)
Observations	320	320

Notes: Dependent variable *withhold evidence* is equal to 1 if the low type agent sent no evidence in the Commitment treatment and 0 if they sent low evidence. *Period* takes values from 1 to 20 and represents the period. *Gender* is a dummy variable that takes the value 1 if subject is female and 0 otherwise. *Risk Aversion* takes the value 1 if the subject is classified as risk averse based on the number of safe options they chose in Activity 1 and 0 otherwise. *Ability to Bayesian update* is a dummy variable that takes the value 1 if subject answered the Activity 2 question of Part II correctly and 0 otherwise. Standard errors are clustered at the individual level. p-values are in parentheses; * p<0.1, ** p<0.05, *** p<0.01.

Table C.5: Probit Regressions Relating Withholding Information to the Difference Between Rewards in the Commitment Treatment Conditioning on the Difference being Positive

	(1)	(2)	(3)
Difference Between	0.022***	0.023***	0.019***
Rewards	(0.000)	(0.000)	(0.000)
Reward for			-0.007**
Low Evidence			(0.017)
Period		0.026*	0.019
		(0.071)	(0.204)
Gender		-0.263	-0.289
		(0.316)	(0.275)
Risk Aversion		-0.848***	-0.871***
		(0.000)	(0.000)
Ability to		0.114	0.067
Bayesian update		(0.669)	(0.799)
Constant	0.046	0.698*	1.155***
	(0.751)	(0.092)	(0.008)
Observations	320	320	320

Notes: Dependent variable *withhold evidence* is equal to 1 if the low type agent sent no evidence in the Commitment treatment and 0 if they sent low evidence. *Difference Between Rewards* is the difference between Reward for No Evidence and Reward for Low Evidence. *Period* takes values from 1 to 20 and represents the period. *Gender* is a dummy variable that takes the value 1 if subject is female and 0 otherwise. *Risk Aversion* takes the value 1 if the subject is classified as risk averse based on the number of safe options they chose in Activity 1 and 0 otherwise. *Ability to Bayesian update* is a dummy variable that takes the value 1 if subject answered the Activity 2 question of Part II correctly and 0 otherwise. Standard errors are clustered at the individual level. p-values are in parentheses; * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

C.3 Model With Guilt

Alternative to the lying aversion model in which the agent was lying averse, we consider a model in which the agent may be guilt averse. Using the simple guilt model of Battigalli and Dufwenberg (2007), a principal who accounts for the agent's guilt aversion solves the following problem:

$$\max_{x_0, x_-} \quad q \cdot (I - (H - x_0)) + (1 - q) \cdot (I - (x_- - L))$$
$$\text{s.t. } x_- \geq x_0 - G \cdot \beta \cdot max\{\xi - (I - (x_0 - L)), 0\}$$

where $G > 0$ is the agent's guilt parameter, $\beta \in [0, 1]$ is the agent's second order belief on the principal's belief that the agent is high type when he sees no-evidence, $\xi \in [I - H, I]$ is the principal's expected payoff when he sees no-evidence.

Using the parameters of the experiment, the problem is:

$$\max_{x_0, x_-} \quad 0.5 \cdot x_0 + 0.5 \cdot (100 - x_-)$$
$$\text{s.t. } x_- \geq x_0 - G \cdot \beta \cdot max\{\xi - (100 - x_0), 0\}$$

In the optimal mechanism: $x_-^C = x_0^C - G \cdot \beta \cdot max\{\xi - (100 - x_0^C), 0\}$, and the principal's maximization problem becomes:

$$\max_{x_0} \quad x_0 + (100 - (x_0 - G \cdot \beta \cdot max\{\xi - (100 - x_0), 0\}))$$

Case I: If $\xi \leq 100 - x_0$

Then, the principal's maximization problem reduces to the model without guilt.

Case II: If $\xi > 100 - x_0$, the principal's maximization problem:

$$\max_{x_0} \quad x_0 + (100 - x_0 + G \cdot \beta \cdot (\xi - 100 + x_0))$$

Fist order condition, $G \cdot \beta$, is strictly increasing in x_0, since $G > 0$ and $\beta \geq 0$. Additionally, $\xi = x_0^C$ in equilibrium. So, optimal rewards are:

$$x_0^C = 100 \text{ , } x_-^C = 100 \cdot (1 - G \cdot \beta) \quad \text{if } G \cdot \beta \leq 1$$

$$x_0^C = 100 \text{ , } x_-^C = 0 \quad \text{if } G \cdot \beta > 1$$

Since we find that the reward for no evidence in the Commitment Treatment, 60.42, is significantly lower than 100 ($p < 0.001$), the simple guilt model does not explain our experimental findings.

C.4 Instructions

C.4.1 Part I Instructions for No-Commitment Treatment

Welcome, and thank you for coming today to participate in this experiment. This is an experiment in decision making. You will receive a $7 participation fee. In addition to that, if you follow the instructions and are careful with your decisions, you can earn a significant amount of money, which will be paid to you privately at the end of the session.

The experiment is expected to finish in 120 minutes. The experiment consists of two independent paying parts and a questionnaire. This is the instructions for Part 1.

In this part of the experiment, you will participate in 20 independent decision periods. At the end of the experiment, the computer will randomly select one decision period for payment. The period selected depends solely upon chance and each period is equally likely. Your final earnings in the experiment will be your earnings in the selected period plus your earnings in Part II and the $7 show-up fee.

Your earnings in this experiment will be calculated in Experimental Currency Units (ECUs). At the end of today's session, all your earnings will be converted to US dollars at a rate of 10 ECUs=$1

During the experiment, it is important that you do not talk to any other subjects. Please turn off your cell phones. If you have a question, please raise your hand, and the experimenter will come by to answer your question. Food or drink is not allowed in the lab; if you have food or drink with you, please keep it stored away in your bags. Failure to comply with these instructions means that you will be asked to leave the experiment and all your earnings will be forfeited.

Instructions

You will be informed of your role as the Sender or the Receiver in the first round of the experiment. Your role will be fixed throughout this part of the experiment. In each period, you will be randomly matched with another subject in this room who will be assigned the other role. There will be a new random matching at the beginning of each period, so you will potentially be matched with different people in different rounds. In each round, the Sender will be randomly assigned a type: High or Low. Each type is equally likely. The value of High type to the Receiver is 100, while the value of the Low type is 0.

The Low type Sender has evidence about their type, while the High type sender doesn't. At the beginning of each round, each Sender will choose a message to send to the Receiver they are matched with in that round. The Low type Sender has a choice between telling the truth or pretending that they don't have evidence. The messages available to the Low type Sender are: "My type is low" and "I don't have evidence for my type". The High type Sender, on the other hand, can only send the message "I don't have evidence for my type". The information is summarized in Table 1.

Type (t)	**Value (v)**	**Probability (p)**	**Available Messages**
High	100	50%	"I don't have evidence for my type"
Low	0	50%	"My type is low", "I don't have evidence for my type"

Table 1

After observing the message that the Sender sent, the Receiver will choose a reward between 0 and 100 to send to the Sender.

Payoffs in Each Round

The Sender's payoff in each round will be equal to the reward chosen by the Receiver for the message the Sender sent.

$$\pi_{Sender} = reward$$

The payoff of the Receiver is:

$$\pi_{Receiver} = 100 - |value - reward|$$

where "value" is the value associated with the Sender's type and "reward" is the reward the Receiver chose for the message the Sender sent. The payoff to the Receiver will be 100 minus the distance between the chosen reward and the value of the Sender. So, the Receiver's ideal point for the reward is equal to the value associated with the Sender's type. Notice that the Receiver can choose any number between 0 and 100 as the reward.

At the end of each round, the Sender's type, the message the Sender chose, and the payoffs of the matched Sender and Receiver will be shown to both players. Then, there will be a new random matching and a new round will begin.

Earnings

Once the experiment is finished, the computer will randomly pick 1 round out of the 20 rounds that you completed. The earnings you made on that round will be your earnings in this part of the experiment. Hence, you should make careful decisions in each round because it might be the paying round.

Questions for Checking Understanding

The first screen in the experiment consists of 2 questions that you need to answer correctly to begin the actual experiment. If you answer any of the questions incorrectly, you will receive a pop-up indicating which question you need to correct. Once you answer both questions correctly, you will be directed to the first period of the experiment.

Are there any questions?

C.4.2 Part I Instructions for Commitment Treatment

Welcome, and thank you for coming today to participate in this experiment. This is an experiment in decision making. You will receive a $7 participation fee. In addition to that, if you follow the instructions and are careful with your decisions, you can earn a significant amount of money, which will be paid to you privately at the end of the session.

The experiment is expected to finish in 120 minutes. The experiment consists of two independent paying parts and a questionnaire. This is the instructions for Part 1.

In this part of the experiment, you will participate in 20 independent decision periods. At the end of the experiment, the computer will randomly select one decision period for payment. The period selected depends solely upon chance and each period is equally likely. Your final earnings in the experiment will be your earnings in the selected period plus your earnings in Part II and the $7 show-up fee.

Your earnings in this experiment will be calculated in Experimental Currency Units (ECUs). At the end of today's session, all your earnings will be converted to US dollars at a rate of 10

ECUs=$1

During the experiment, it is important that you do not talk to any other subjects. Please turn off your cell phones. If you have a question, please raise your hand, and the experimenter will come by to answer your question. Food or drink is not allowed in the lab; if you have food or drink with you, please keep it stored away in your bags. Failure to comply with these instructions means that you will be asked to leave the experiment and all your earnings will be forfeited.

Instructions

You will be informed of your role as the Sender or the Receiver in the first round of the experiment. Your role will be fixed throughout this part of the experiment. In each period, you will be randomly matched with another subject in this room who will be assigned the other role. There will be a new random matching at the beginning of each period, so you will potentially be matched with different people in different rounds.

In each round, the Sender will be randomly assigned a type: High or Low. Each type is equally likely. The value of High type to the Receiver is 100, while the value of the Low type is 0.

At the beginning of each round, the Receiver will choose a reward between 0 and 100 for each message that they can possibly receive. After observing the reward scheme, the Sender will choose which message to send.

The Low type Sender has evidence about their type, while the High type sender doesn't. After observing the reward scheme, each Sender will choose a message to send to the Receiver they are matched with in that round. The Low type Sender has a choice between telling the truth or pretending that they don't have evidence. The messages available to the Low type Sender

are: “My type is low” and “I don’t have evidence for my type”. The High type Sender, on the other hand, can only send the message “I don’t have evidence for my type”. The information is summarized in Table 1.

Type (t)	Value (v)	Probability (p)	Available Messages
High	100	50%	“I don’t have evidence for my type”
Low	0	50%	“My type is low”, “I don’t have evidence for my type”

Table 1

Payoffs in Each Round

The Sender’s payoff in each round will be equal to the reward chosen by the Receiver for the message the Sender sent.

$$\pi_{Sender} = reward$$

The payoff of the Receiver is:

$$\pi_{Receiver} = 100 - |value - reward|$$

where ”value” is the value associated with the Sender’s type and ”reward” is the reward the Receiver chose for the message the Sender sent. The payoff to the Receiver will be 100 minus the distance between the chosen reward and the value of the Sender. So, the Receiver’s ideal point for the reward is equal to the value associated with the Sender’s type. Notice that the Receiver can choose any number between 0 and 100 as the reward. At the end of each round, the Sender’s type, the message the Sender chose, and the payoffs of the matched Sender and Receiver will be shown to both players. Then, there will be a new random matching and a new round will begin.

Earnings

Once the experiment is finished, the computer will randomly pick 1 round out of the 20 rounds that you completed. The earnings you made on that round will be your earnings in this part of the experiment. Hence, you should make careful decisions in each round because it might be the paying round.

Questions for Checking Understanding

The first screen in the experiment consists of 2 questions that you need to answer correctly to begin the actual experiment. If you answer any of the questions incorrectly, you will receive a pop-up indicating which question you need to correct. Once you answer both questions correctly, you will be directed to the first period of the experiment.

Are there any questions?

C.4.3 Screenshots from the Experiment

Figure C.1: Screen of a High Type Sender, No-Commitment Treatment

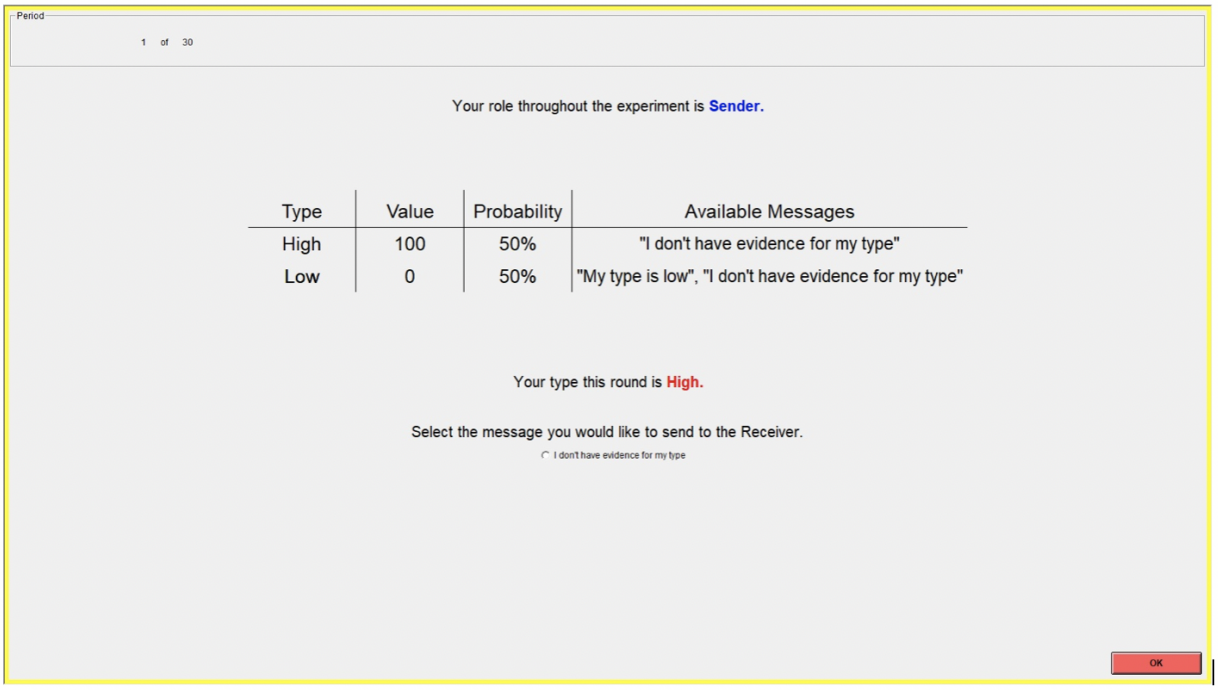


Figure C.2: Screen of a Low Type Sender, No-Commitment Treatment

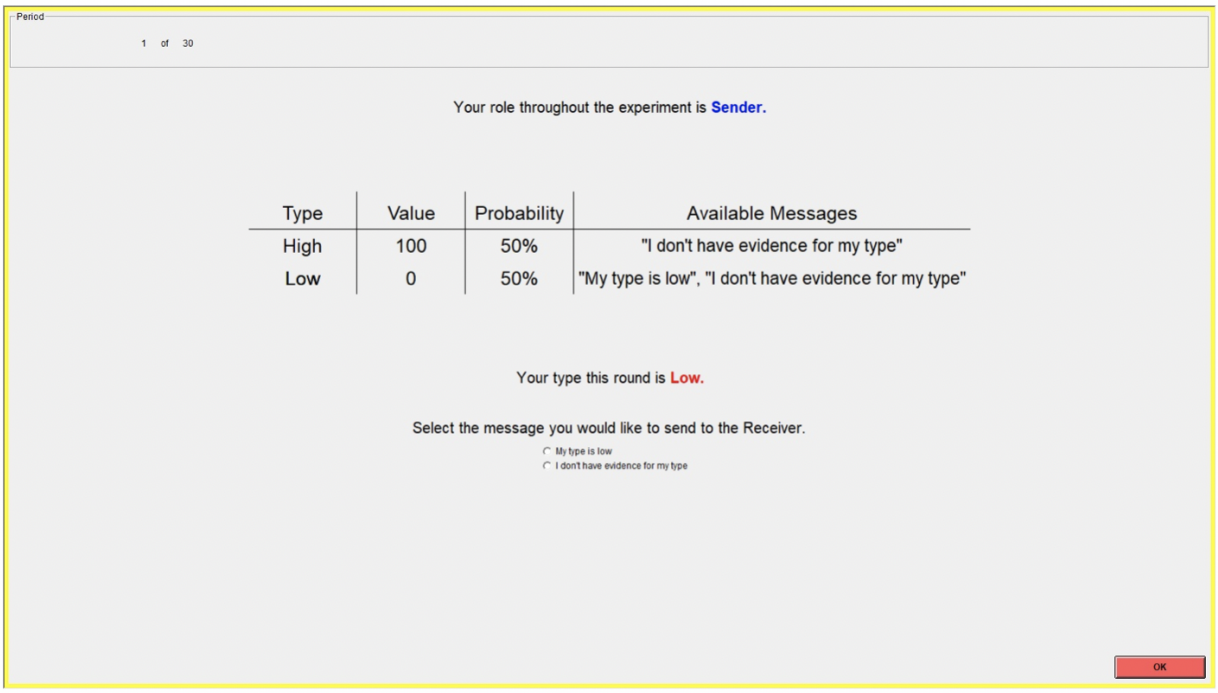
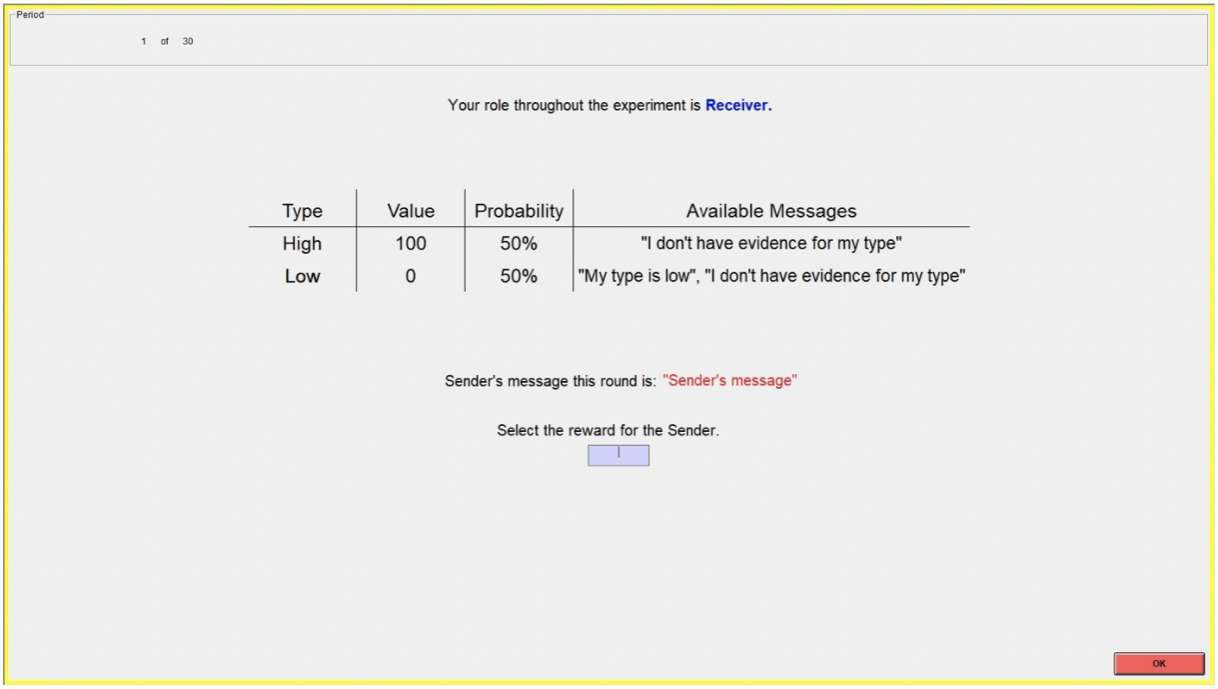


Figure C.3: Screen of a Receiver, No-Commitment Treatment



Notes: The message in the real experiment was either "My type is low" or "I don't have evidence for my type" based on the Sender's choice.

Figure C.4: Screen of a Receiver, Commitment Treatment

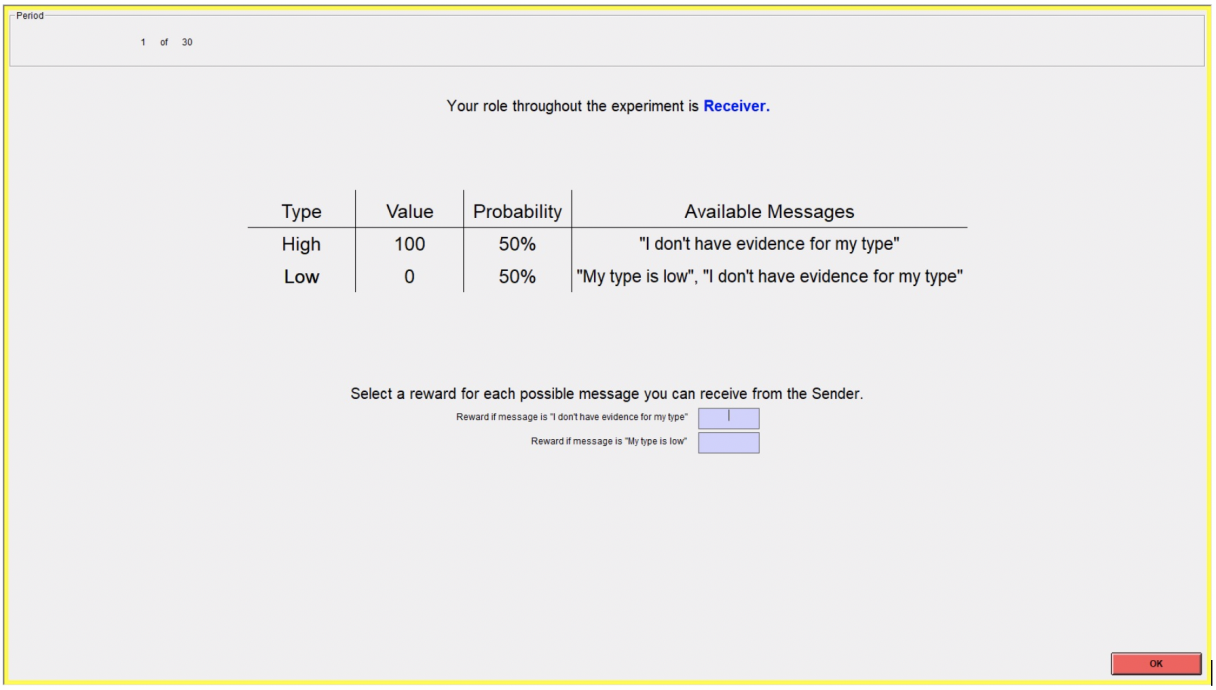
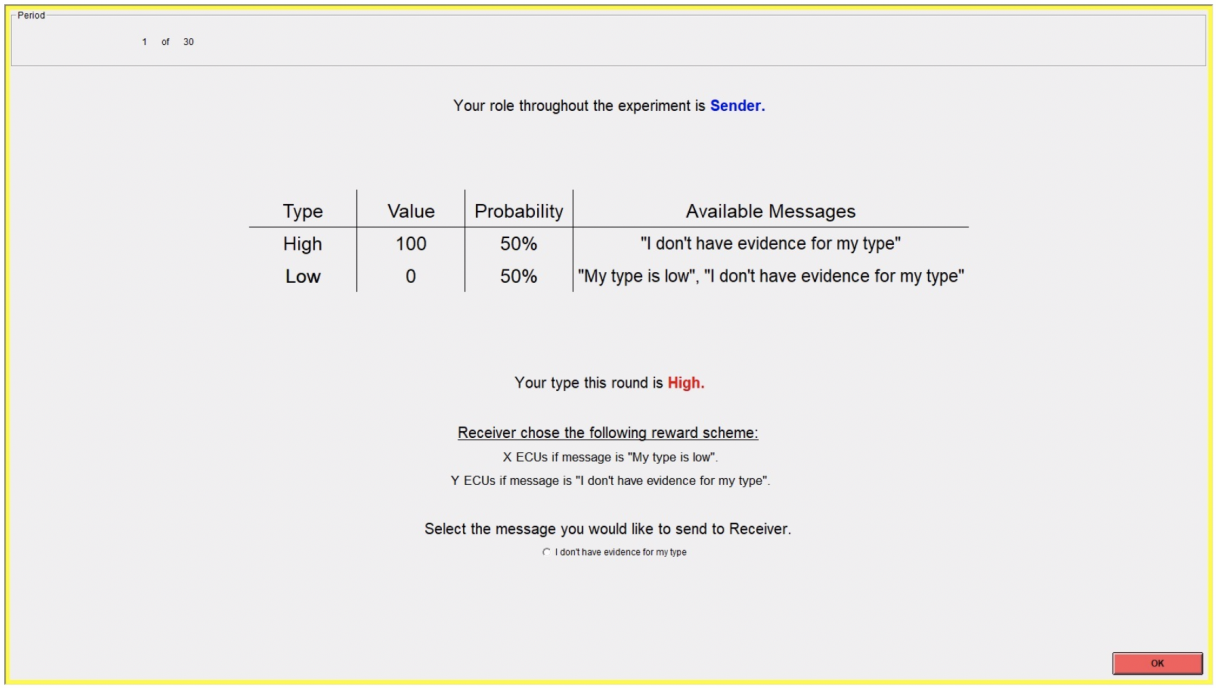
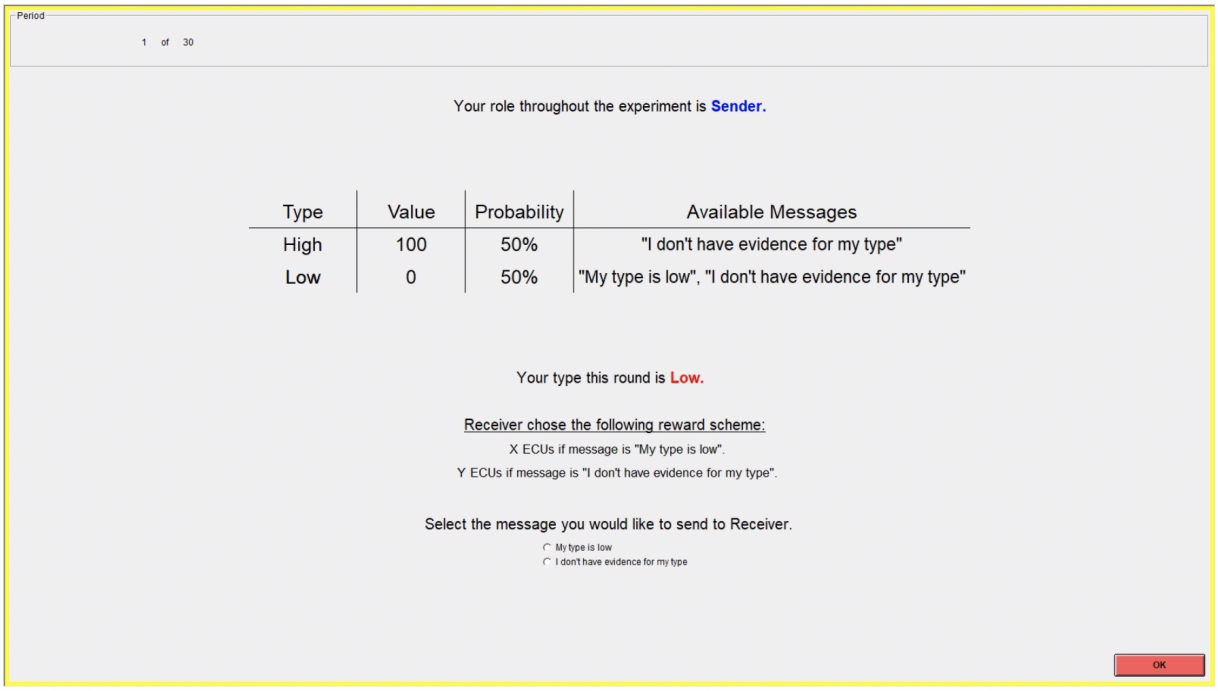


Figure C.5: Screen of a High Type Sender, Commitment Treatment



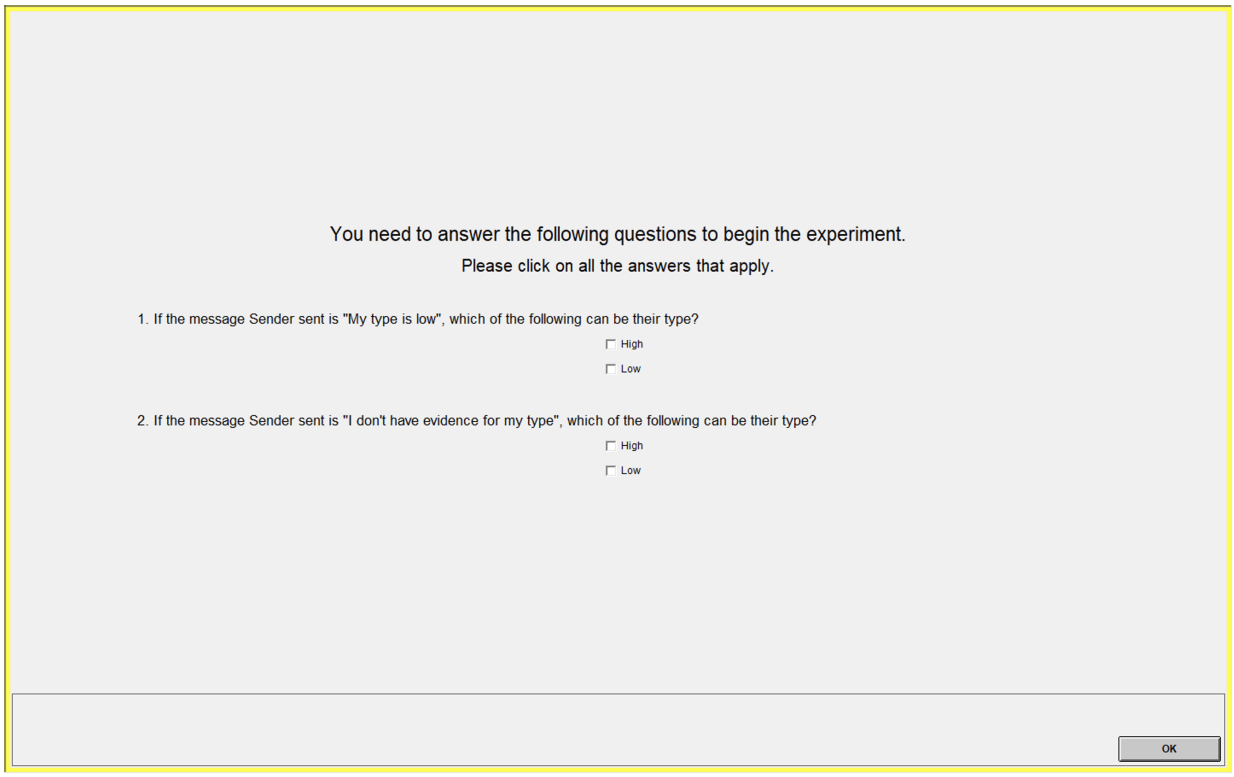
Notes: The rewards in the real experiment were numbers between 0 and 100 that the Receiver chose.

Figure C.6: Screen of a Low Type Sender, Commitment Treatment



Notes: The rewards in the real experiment were numbers between 0 and 100 that the Receiver chose.

Figure C.7: Questions for Checking Understanding



Bibliography

J. Abeler, D. Nosenzo, and C. Raymond. Preferences for truth-telling. *Econometrica*, 87(4): 1115–1153, 2019.

G. Azmat, M. Bagues, A. Cabrales, and N. Iriberri. What you don't know… can't hurt you? a natural field experiment on relative performance feedback in higher education. *Management Science*, 65(8):3714–3736, 2019.

L. Babcock, M. P. Recalde, L. Vesterlund, and L. Weingart. Gender differences in accepting and receiving requests for tasks with low promotability. *American Economic Review*, 107(3): 714–47, 2017.

B. M. Barber and T. Odean. Boys will be boys: Gender, overconfidence, and common stock investment. *The quarterly journal of economics*, 116(1):261–292, 2001.

K. Barron. Belief updating: does the 'good-news, bad-news' asymmetry extend to purely financial domains? *Experimental Economics*, 24(1):31–58, 2021.

B. Bartling and U. Fischbacher. Shifting the blame: On delegation and responsibility. *The Review of Economic Studies*, 79(1):67–87, 2012.

P. Battigalli and M. Dufwenberg. Guilt in games. *American Economic Review*, 97(2):170–176, 2007.

P. Battigalli, G. Charness, and M. Dufwenberg. Deception: The role of guilt. *Journal of Economic Behavior & Organization*, 93:227–232, 2013.

E. Ben-Porath, E. Dekel, and B. L. Lipman. Mechanisms with evidence: Commitment and rbustness. *Econometrica*, 87(2):529–566, 2019.

D. J. Benjamin. Errors in probabilistic reasoning and judgment biases. *Handbook of Behavioral Economics: Applications and Foundations 1*, 2:69–186, 2019.

S. Beyer. Gender differences in the accuracy of self-evaluations of performance. *Journal of personality and social psychology*, 59(5):960, 1990.

J. A. Bohren, A. Imas, and M. Rosenberg. The dynamics of discrimination: Theory and evidence. *American economic review*, 109(10):3395–3436, 2019.

A. Born, E. Ranehill, and A. Sandberg. Gender and willingness to lead: Does the gender composition of teams matter? *The Review of Economics and Statistics*, pages 1–46, 2020.

J. Brandts and G. Charness. The strategy versus the direct-response method: a first survey of experimental comparisons. *Experimental Economics*, 14(3):375–398, 2011.

J. Brandts and C. Rott. Advice from women and men and selection into competition. *Journal of Economic Psychology*, 82:102333, 2021.

J. Brandts, V. Groenert, and C. Rott. The impact of advice on women's and men's selection into competition. *Management Science*, 61(5):1018–1035, 2015.

J. Bull and J. Watson. Hard evidence and mechanism design. *Games and Economic Behavior*, 58(1):75 – 93, 2007. ISSN 0899-8256. doi: https://doi.org/10.1016/j.geb.2006.03.003. URL http://www.sciencedirect.com/science/article/pii/S0899825606000352.

T. Buser, L. Gerhards, and J. Van Der Weele. Responsiveness to feedback as a personal trait. *Journal of Risk and Uncertainty*, 56(2):165–192, 2018.

A. C. Cameron and D. L. Miller. A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 50(2):317–372, 2015.

A. C. Cameron, J. B. Gelbach, and D. L. Miller. Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics*, 90(3):414–427, 08 2008.

P. Catalyst. Women in s&p 500 companies. *Available at SSRN 3239814*, 2020.

B. Çelen, S. Kariv, and A. Schotter. An experimental test of advice and social learning. *Management Science*, 56(10):1687–1701, 2010.

P. Chakraborty, D. Serra, et al. Gender and leadership in organizations: Promotions, demotions and angry workers. *Working Papers 20210104–001*, 2021.

G. Charness and C. Dave. Confirmation bias with motivated beliefs. *Games and Economic Behavior*, 104:1–23, 2017.

G. Charness and M. Dufwenberg. Promises and partnership. *Econometrica*, 74(6):1579–1601, 2006.

G. Charness and D. Levin. When optimal choices feel wrong: A laboratory study of bayesian updating, complexity, and affect. *American Economic Review*, 95(4):1300–1309, September 2005. doi: 10.1257/0002828054825583. URL https://www.aeaweb.org/articles?id=10.1257/0002828054825583.

A. Chaudhuri, A. Schotter, and B. Sopher. Talking ourselves to efficiency: Coordination in intergenerational minimum effort games with private, almost common and common knowledge of advice. *The Economic Journal*, 119(534):91–122, 2009.

D. L. Chen, M. Schonger, and C. Wickens. otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97, 2016.

J. Chen and D. Houser. When are women willing to lead? the effect of team gender composition and gendered tasks. *The Leadership Quarterly*, 30(6):101340, 2019.

K. Coffman, M. Collis, and L. Kulkarni. *Stereotypes and belief updating*. Harvard Business School, 2019.

K. Coffman, C. B. Flikkema, and O. Shurchkov. Gender stereotypes in deliberation and team decisions. *Games and Economic Behavior*, 2021a.

K. B. Coffman. Evidence on self-stereotyping and the contribution of ideas. *The Quarterly Journal of Economics*, 129(4):1625–1660, 2014.

K. B. Coffman, P. U. Araya, and B. Zafar. A (dynamic) investigation of stereotypes, belief-updating, and behavior. 2021b.

K. B. Coffman, C. L. Exley, and M. Niederle. The role of beliefs in driving gender discrimination. *Management Science*, 2021c.

D. J. Cooper and J. H. Kagel. A failure to communicate: an experimental investigation of the effects of advice on strategic play. *European Economic Review*, 82:24–45, 2016.

P. Cortés, J. Pan, L. Pilossoph, and B. Zafar. Gender differences in job search and the earnings gap: Evidence from business majors. 2021.

A. Coutts. Good news and bad news are still news: Experimental evidence on belief updating. *Experimental Economics*, 22(2):369–395, 2019.

A. Coutts, L. Gerhards, and Z. Murad. What to blame? self-serving attribution bias with multi-dimensional uncertainty. 2020.

V. P. Crawford and J. Sobel. Strategic information transmission. *Econometrica*, pages 1431–1451, 1982.

D. Danz, L. Vesterlund, and A. J. Wilson. Belief elicitation: Limiting truth telling with information on incentives. 2020.

R. Deneckere and S. Severinov. Mechanism design with partial state verifiability. *Games and Economic Behavior*, 64(2):487–513, 2008.

R. Deneckere and S. Severinov. Screening, signalling and costly misrepresentation. Technical report, Working paper, 2017.

G. Di Bartolomeo, D. Martin, S. Papa, R. Laura, et al. Guilt aversion: Eve versus adam. In *Guilt Aversion: Eve versus Adam*. 2022.

T. Ding and A. Schotter. Matching and chatting: An experimental study of the impact of network communication on school-matching mechanisms. *Games and Economic Behavior*, 103:94–115, 2017.

T. Ding and A. Schotter. Learning and mechanism design: An experimental test of school matching mechanisms with intergenerational advice. *The Economic Journal*, 129(623):2779–2804, 2019.

C. Drobner. Motivated beliefs and anticipation of uncertainty resolution. *American Economic Review: Insights*, 4(1):89–105, 2022.

R. A. Dye. Disclosure of nonproprietary information. *Journal of Accounting Research*, pages 123–145, 1985.

C. C. Eckel and P. J. Grossman. Men, women and risk aversion: Experimental evidence. *Handbook of experimental economics results*, 1:1061–1073, 2008.

F. Ederer and E. Fehr. Deception and incentives: How dishonesty undermines effort provision. Technical report, Working paper, 2017.

W. Edwards. Conservatism in human information processing. *Formal representation of human judgment*, 1968.

D. Eil and J. M. Rao. The good news-bad news effect: asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, 3(2):114–38, 2011.

S. Erat. Avoiding lying: The case of delegated deception. *Journal of Economic Behavior & Organization*, 93:273–278, 2013.

S. Erat and U. Gneezy. White lies. *Management Science*, 58(4):723–733, 2012.

S. Ertac. Does self-relevance affect information processing? experimental evidence on the response to performance and non-performance feedback. *Journal of Economic Behavior & Organization*, 80(3):532–545, 2011.

S. Ertac and M. Y. Gurdal. Deciding to decide: Gender, leadership and risk-taking in groups. *Journal of Economic Behavior & Organization*, 83(1):24–30, 2012.

S. Ertac, L. Koçkesen, and D. Ozdemir. The role of verifiability and privacy in the strategic provision of performance feedback: Theory and experimental evidence. *Games and Economic Behavior*, 100:24–45, 2016.

L. Eskreis-Winkler, K. L. Milkman, D. M. Gromet, and A. L. Duckworth. A large-scale field experiment shows giving advice improves academic outcomes for the advisor. *Proceedings of the national academy of sciences*, 116(30):14808–14810, 2019.

C. L. Exley and J. B. Kessler. The gender gap in self-promotion. Technical report, National Bureau of Economic Research, 2019.

U. Fischbacher. z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2):171–178, 2007.

U. Fischbacher and F. Föllmi-Heusi. Lies in disguise—an experimental study on cheating. *Journal of the European Economic Association*, 11(3):525–547, 2013.

G. R. Fréchette, A. Lizzeri, and J. Perego. Rules and commitment in communication. *CEPR Discussion Paper No. DP14085*, 2019.

S. Frederick. Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19 (4):25–42, 2005.

D. Friedman. Monty hall's three doors: Construction and deconstruction of a choice anomaly. *American Economic Review*, 88(4):933–946, 1998.

L. Friesen and L. Gangadharan. Designing self-reporting regimes to encourage truth telling: An experimental study. *Journal of Economic Behavior & Organization*, 94:90–102, 2013.

R. Fry. Women now outnumber men in the u.s. college-educated labor force. 2022.

J. Glazer and A. Rubinstein. A study in the pragmatics of persuasion: a game theoretical approach. *Theoretical Economics*, 1(4):395–410, 2006.

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

U. Gneezy and J. Potters. An experiment on risk taking and evaluation periods. *The quarterly journal of economics*, 112(2):631–645, 1997.

U. Gneezy, A. Kajackaite, and J. Sobel. Lying aversion and the size of the lie. *American Economic Review*, 108(2):419–53, 2018.

U. Gneezy, S. Saccardo, M. Serra-Garcia, and R. van Veldhuizen. Bribing the self. *Games and Economic Behavior*, 120:311–324, 2020.

E. Goldman and S. L. Slezak. An equilibrium model of incentive contracts in the presence of information manipulation. *Journal of Financial Economics*, 80(3):603–626, 2006.

J. R. Green and J.-J. Laffont. Partially verifiable information and mechanism design. *Review of Economic Studies*, 53(3):447–456, 1986.

B. Greiner. Subject pool recruitment procedures: organizing experiments with orsee. *Journal of the Economic Science Association*, 1(1):114–125, 2015.

D. M. Grether. Bayes rule as a descriptive model: The representativeness heuristic. *The Quarterly journal of economics*, 95(3):537–557, 1980.

D. M. Grether. Testing bayes rule and the representativeness heuristic: Some experimental evidence. *Journal of Economic Behavior & Organization*, 17(1):31–57, 1992.

S. J. Grossman. The informational role of warranties and private disclosure about product quality. *Journal of Law and Economics*, 24(3):461–483, 1981.

S. J. Grossman and O. D. Hart. Disclosure laws and takeover bids. *Journal of Finance*, 35(2): 323–334, 1980.

Z. Grossman and D. Owens. An unlucky feeling: Overconfidence and noisy feedback. *Journal of Economic Behavior & Organization*, 84(2):510–524, 2012.

I. Guttman, O. Kadan, and E. Kandel. A rational expectations theory of kinks in financial reporting. *Accounting Review*, 81(4):811–848, 2006.

S. Hart, I. Kremer, and M. Perry. Evidence games: Truth and commitment. *American Economic Review*, 107(3):690–713, March 2017. doi: 10.1257/aer.20150913. URL `https://www.aeaweb.org/articles?id=10.1257/aer.20150913`.

E. Heikensten and S. Isaksson. Simon says: Examining gender differences in advice seeking and influence in the lab. *Available at SSRN 3273186*, 2019.

M. Hinnosaar. Gender inequality in new media: Evidence from wikipedia. *Journal of Economic Behavior & Organization*, 163:262–276, 2019.

C. A. Holt and S. K. Laury. Risk aversion and incentive effects. *American Economic Review*, 92 (5):1644–1655, 2002.

C. A. Holt and A. M. Smith. An update on bayesian updating. *Journal of Economic Behavior & Organization*, 69(2):125–134, 2009.

R. Iyengar and A. Schotter. Learning under supervision: an experimental study. *Experimental Economics*, 11(2):154–173, 2008.

G. Z. Jin, M. Luca, and D. Martin. Is no news (perceived as) bad news? an experimental investigation of information disclosure. *American Economic Journal: Microeconomics*, 13(2): 141–73, 2021.

D. Kahneman and A. Tversky. Subjective probability: A judgment of representativeness. *Cognitive psychology*, 3(3):430–454, 1972.

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

N. Kartik. Strategic communication with lying costs. *Review of Economic Studies*, 76(4):1359–1395, 2009.

N. Kartik, M. Ottaviani, and F. Squintani. Credulity, lies, and costly talk. *Journal of Economic theory*, 134(1):93–116, 2007.

P. Kline and A. Santos. A score based approach to wild bootstrap inference. *Journal of Econometric Methods*, 1(1):23–41, 2012.

B. Köszegi. Ego utility, overconfidence, and task choice. *Journal of the European Economic Association*, 4(4):673–707, 2006.

J. M. Lacker and J. A. Weinberg. Optimal contracts under costly state falsification. *Journal of Political Economy*, 97(6):1345–1363, 1989.

A. Landier. Wishful thinking: A model of optimal reality denial. *Massachusetts Institute*, 2000.

E. P. Lazear and S. Rosen. Male-female wage differentials in job ladders. *Journal of Labor Economics*, 8(1, Part 2):S106–S123, 1990.

S. Manian and K. Sheth. Follow my lead: Assertive cheap talk and the gender gap. *Management Science*, 2021.

G. Mayraz. Priors and desires–a model of payoff-dependent beliefs. 2009.

P. R. Milgrom. Good news and bad news: Representation theorems and applications. *Bell Journal of Economics*, pages 380–391, 1981.

M. M. Möbius, M. Niederle, P. Niehaus, and T. S. Rosenblat. Managing self-confidence: Theory and experimental evidence. *Management Science*, 2022.

R. Newson et al. somersd-confidence intervals for nonparametric statistics and their differences. *Stata Technical Bulletin*, 10(55), 2001.

M. Niederle and L. Vesterlund. Do women shy away from competition? do men compete too much? *The quarterly journal of economics*, 122(3):1067–1101, 2007.

T. Nihonsugi, T. Tanaka, and M. Haruno. Gender differences in guilt aversion in korea and the united kingdom. *Scientific Reports*, 12(1):8187, 2022.

Y. Nyarko, A. Schotter, and B. Sopher. On the informational content of advice: A theoretical and experimental study. *Economic Theory*, 29(2):433–452, 2006.

R. Peeters and M. Vorsatz. Simple guilt and cooperation. *Journal of Economic Psychology*, 82: 102347, 2021.

M. Rabin and J. L. Schrag. First impressions matter: A model of confirmatory bias. *The quarterly journal of economics*, 114(1):37–82, 1999.

D. Roodman, M. Ø. Nielsen, J. G. MacKinnon, and M. D. Webb. Fast and wild: Bootstrap inference in stata using boottest. *The Stata Journal*, 19(1):4–60, 2019.

D. Rothenhäusler, N. Schweizer, and N. Szech. Guilt in voting and public good games. *European Economic Review*, 101:664–681, 2018.

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

S. Sánchez-Pagés and M. Vorsatz. Enjoy the silence: an experiment on truth-telling. *Experimental Economics*, 12(2):220–241, 2009.

A. Schotter. Decision making with naive advice. *American Economic Review*, 93(2):196–201, 2003.

P. Schwardmann and J. Van der Weele. Deception and self-deception. *Nature human behaviour*, 3(10):1055–1061, 2019.

R. Selten. Axiomatic characterization of the quadratic scoring rule. *Experimental Economics*, 1 (1):43–61, 1998.

M. Serra-Garcia, E. Van Damme, and J. Potters. Hiding an inconvenient truth: Lies and vagueness. *Games and Economic Behavior*, 73(1):244–261, 2011.

M. Serra-Garcia, E. Van Damme, and J. Potters. Lying about what you know or about what you do? *Journal of the European Economic Association*, 11(5):1204–1229, 2013.

G. K. Shastry, O. Shurchkov, and L. L. Xia. Luck or skill: How women and men react to noisy feedback. *Journal of Behavioral and Experimental Economics*, 88:101592, 2020.

I. Sher. Credibility and determinism in a game of persuasion. *Games and Economic Behavior*, 71(2):409, 2011.

E. A. Shrider, M. Kollar, F. Chen, J. Semega, et al. Income and poverty in the united states: 2020. *US Census Bureau, Current Population Reports*, (P60-273), 2021.

O. Svenson. Are we all less risky and more skillful than our fellow drivers? *Acta psychologica*, 47(2):143–148, 1981.

H. Zhou and A. Fishbach. The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of personality and social psychology*, 111(4):493, 2016.

F. Zimmermann. The dynamics of motivated beliefs. *American Economic Review*, 110(2):337–61, 2020.