




Gender differences in advice giving

Elif B. Osun¹ 

Received: 3 March 2022 / Revised: 10 September 2024 / Accepted: 13 September 2024 /

Published online: 15 November 2024

© The Author(s), under exclusive licence to Economic Science Association 2024

Abstract

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. Introducing a model of guilt and responsibility, I discuss possible underlying mechanisms that are consistent with the findings.

Keywords Gender differences · Advice giving · Group decision making · Experiments

JEL Codes C72 · J16 · C90

I am extremely grateful to Emel Filiz-Ozbay for her time, encouragement, and advice on this project. I am thankful to Billur Aksoy, Ian Chadd, Katherine B. Coffman, Keaton Ellis, Seda Ertac, Yusufcan Masatlioglu, Yesim Orhun, Erkut Y. Ozbay, John Shea, Neslihan Uler, and Ece Yegane for their very helpful comments and suggestions. I also thank the editor, the co-editor, and two anonymous reviewers for their very constructive and helpful input. I gratefully acknowledge the research support provided by Behavioral and Social Sciences Dean's Research Initiative Award, University of Maryland, College Park. I would like to express my sincere appreciation to Bates White for the flexibility provided during the course of editing this paper. The opinions expressed represent only those of the author, and do not represent the views or opinions of Bates White, LLC or of other Bates White employees or affiliates.

The replication material for the study is available at <https://doi.org/10.17605/OSF.IO/SDFBE>.

✉ Elif B. Osun
elif.osun@bateswhite.com

¹ Bates White Economic Consulting, Washington DC, USA

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 more reluctant to give advice compared to men, this could contribute to there being 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.

Several factors 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 example, Lazear and Rosen 1990, Bohren et al. 2019, Coffman et al. 2021b). Aside from discrimination, underlying preference differences may also be responsible for there being fewer women in higher-profile positions. For example, women are shown to be less competitive than men (Niederle & 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 & 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 in giving advice, and (iii) investigates the effect of the enforceability of advice on the decision to send it. The first point aims to identify whether such a gender gap exists, and the latter two aim to provide insight into 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 and that the gender gap in willingness to send advice is closely related to question difficulty. The gender gap is largest and significant for difficult questions, not significant for medium questions, and even reverses in easy questions, with women sending more advice than men. The results indicate that women are not always reluctant to give advice, but tend to shy away from giving advice as the task becomes harder. 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. Furthermore, I show that the gender gap disappears in a group decision-making setup in which advice becomes enforceable.

The experimental variation in the difficulty of the task and the enforceability of advice provides additional insights into the mechanisms contributing to the gender difference in advice giving. To my knowledge, this is the first paper

systematically analyzing the relationship between the difficulty of a task and the propensity to give advice, and showing that a gender gap in advice-giving arises as question difficulty increases. This is also the first paper to show that the gender differential in advice giving is affected by whether advice is enforceable. The experimental design in this paper cannot identify the mechanism leading to the documented gender gap, but proposes several candidate mechanisms. I introduce a model of guilt and responsibility to propose candidate mechanisms responsible for the observed gender gap. This model is aimed to guide future research in identifying these mechanisms.

The first objective of this study is to determine if there is a fundamental difference in the 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 to make the experimental design more complex to mimic a real-life manager position (such as accounting for the 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.

The remainder of this paper is organized as follows. Section 2 discusses the related literature. Section 3 presents the experimental design and procedures. Section 4 introduces a simple model of guilt and responsibility, presents the model's predictions, and introduces main hypotheses of the experiment. Section 5 explores the nature of the task. Section 6 presents the experimental results. Section 7 conducts robustness analysis. Section 8 discusses the predictions of the model of guilt and responsibility in light of the experimental findings and insights from the gender literature. Section 9 concludes.

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 & Houser, 2019), and gender incongruity 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 & Schotter, 2008; Çelen et al., 2010), influence selection into competition (Brandts et al., 2015), help subjects with strategic play (Cooper & Kagel, 2016), and affect truthful revelation in school matching mechanisms (Ding & 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.

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 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 Fig. 1.

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 1 depicts the

¹ See Appendix B.6 for the exact wording of comprehension questions.

² The exact questions and their order can be found in Table A.1.

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 s 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 Sect. 5 and I report additional robustness analyses around the cutoffs in Sect. 7.

Each image stayed on the screen for 10 s, 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 s. 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 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

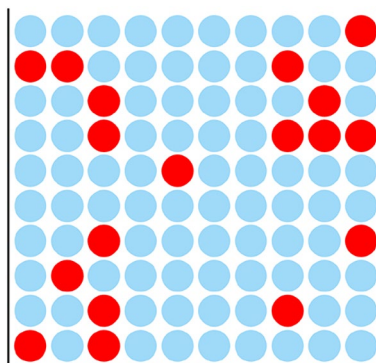
³ For example, I consider a box that contains 95 red balls as easy because the subjects can easily count the number of blue balls (5) and reach the correct answer (95) by subtracting the number of blue balls from 100.

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.

Fig. 1 Box containing a mix of 100 red and blue balls



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 h 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. So, pairs were paid at the pair-level, never individually, regardless of whether the sender sent their guess or not. 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.

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

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 A.2.

Table 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

treatment are summarized below (see Appendix B.2-B.5 for the instructions provided to the subjects).

Advice Treatment In each round, the senders saw the box on their screen for 10 s. 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 s, 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 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.

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

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 Gneezy and Potters (1997), in which the subjects chose how much of their endowment to allocate between a safe and a risky option (see Appendix B.7 and Appendix B.8 for the screenshots of these two tasks).

4 A model of guilt and responsibility

This section introduces a model of guilt and responsibility. The experimental findings, reported in Sect. 6 of this paper, document gender differences in advice giving and examine and rule out some potential mechanisms. However, the exact mechanism behind gender differences in advice giving cannot be identified within the scope of this paper. The purpose of this section is to guide the exploration by outlining several potential underlying mechanisms that may explain the observed findings.

Starting with Charness and Dufwenberg (2006) and Battigalli and Dufwenberg (2007), guilt aversion has been widely studied in various settings. Guilt aversion has been shown to be relevant in settings such as deception (Battigalli et al., 2013), voting and public good games (Rothenhäusler et al. (2018), and cooperation (Peeters & Vorsatz, 2021). Guilt is a potential mechanism affecting subjects' behavior in this experiment; subjects may experience guilt when sending a misleading guess. Although 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 did not anticipate making error-free guesses in these questions.

The intuition behind 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 introduced in this section, an agent's guilt sensitivity is related to how responsible they feel for others' payoffs. In much 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, thereby determining both their and the principal's earnings. On the other hand, in cases where the outcomes of subjects are determined by 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 that assesses how responsible a subject is for 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 for their pair's earnings affects their guilt sensitivity.

Subsections 4.1 and 4.2 relate the model to the Advice and Dictator Treatments, respectively. Subsection 4.3 examines the effect of each parameter on model predictions at the subject level. Subsection 4.4 discusses the effect of question difficulty on model parameters at the question level.

4.1 Advice treatment

Consider a sender S whose guess for the number of red balls in the box in question i is $x_{S,i}$ and a receiver R whose guess without advice is (or would have been) $x_{R,i}$. Let the subjects' expected earnings associated with answers $x_{S,i}$ and $x_{R,i}$ be $m_{S,i}$ and $m_{R,i}$, 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 pair. Denote the sender's expectation of the answer that gets submitted as the pair's final answer as $x_{SR,i}$ in the case that sender sends advice such that $x_{SR,i} = \beta_{S,i}x_{S,i} + (1 - \beta_{S,i})x_{R,i}$, where $\beta_{S,i} \in [0, 1]$. $\beta_{S,i}$ is the belief of the sender on how influential their answer is. In the extremes, if $\beta_{S,i} = 0$, the sender believes that their answer has no affect on the receiver's final answer, and if $\beta_{S,i} = 1$, the sender believes that the receiver follows the sender's advice verbatim. Given the pay-off structure used in the experiment, the payoff associated with the answer will be $m_{SR,i} \in [\min\{m_{R,i}, m_{S,i}\}, \max\{m_{R,i}, m_{S,i}\}]$.

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 pair'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 pair's final answer if they choose to send their guess. In the Advice Treatment, $\rho = \alpha$, while in the Dictator Treatment, $\rho = 1$. Note that $1 - \rho$ can be thought of as the scope for shifting the blame for the pair'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, $G_{S,i}(a_i; \rho, d_i)$ as follows:

$$G_{S,i}(a_i; \rho, d_i) = \begin{cases} 0 & \text{if } a_i = \text{not and } d_i = R \\ w_{S,i} \times \Gamma_{S,i}(\rho) & \text{if } a_i = \text{send and } d_i = R \\ w_{S,i} & \text{if } d_i = S \end{cases} \quad (1)$$

where $a_i \in \{\text{send}, \text{not}\}$ is the action of the sender in question i , $w_{S,i}$ 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, $d_i \in \{S, R\}$ denotes the role of the subject

whose answer gets implemented as the pair's final answer in question i , $\Gamma_{S_i}(\rho) \rightarrow [0, 1] \times [0, 1]$ and $\frac{d\Gamma_{S_i}(\rho)}{d\rho} \geq 0$.¹⁰

The functional form of $G_{S_i}(a_i; \rho, d_i)$ 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.

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. With this assumption, the expected utility of a sender S in the Advice Treatment upon sending their guess in question i is:

$$u_{S,i,send}^A(m_{S,i}, m_{R,i}, \alpha, \beta_{S,i}) = (1 - \alpha) \times [m_{SR,i}(\beta_{S,i}) - w_{S,i} \times \Gamma_{S,i}(\alpha) \times (m_{max} - m_{SR,i}(\beta_{S,i}))] \\ + \alpha \times [m_{S,i} - w_{S,i} \times (m_{max} - m_{S,i})] \quad (2)$$

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

The expected utility of a sender S in the Advice Treatment if they do not send their guess in question i is represented by:

$$u_{S,i,not}^A(m_{S,i}, m_{R,i}, \alpha, \beta_{S,i}) = (1 - \alpha) \times m_{R,i} + \alpha \times [m_{S,i} - w_{S,i} \times (m_{max} - m_{S,i})] \quad (3)$$

where the first term corresponds to the case in which the receiver's guess is implemented as the final answer of the pair, so the sender expects the pair's monetary earnings to be $m_{R,i}$, 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, so the sender expects the pair's monetary earnings to be $m_{S,i}$, feels full responsibility for

¹⁰ In a sense, $\Gamma_{S_i}(\rho)$ can be thought of as a weight on the guilt intensity parameter, and the senders' guilt intensity is 'discounted' if they shift some of the blame to the receiver. Note that there is no room to shift the blame in the Dictator Treatment, so the guilt intensity parameter receives the maximum weight of 1 for all senders.

the pair's earnings since their answer is implemented as the pair's final answer so their guilt sensitivity is $w_{S,i}$, and expects to disappoint the receiver by $m_{max} - m_{S,i}$.

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

$$\frac{m_{SR,i}(\beta_{S,i}) - m_{R,i}}{m_{max} - m_{SR,i}(\beta_{S,i})} \geq w_{S,i} \times \Gamma_{S,i}(\alpha) \quad (4)$$

4.2 Dictator treatment

In the Dictator Treatment, when a sender sends their guess $x_{S,i}$, it is implemented as the final answer of the pair. Hence, there is no room for beliefs on how influential the sender believes they are.¹¹ The expected utility of a sender S in the Dictator Treatment upon sending their guess in question i :

$$u_{S,i,send}^D(m_{S,i}, m_{R,i}, \alpha, \beta_{S,i}) = m_{S,i} - w_{S,i} \times (m_{max} - m_{S,i}) \quad (5)$$

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

The expected utility of a sender S in the Dictator Treatment if they do not send their guess in question i is represented by:

$$u_{S,i,not}^D(m_{S,i}, m_{R,i}, \alpha, \beta_{S,i}) = (1 - \alpha) \times m_{R,i} + \alpha \times [m_{S,i} - w_{S,i} \times (m_{max} - m_{S,i})] \quad (6)$$

where the first term corresponds to the case in which the receiver's guess is implemented as the final answer of the pair, so the sender expects the pair's monetary earnings to be $m_{R,i}$, 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, so the sender expects the pair's monetary earnings to be $m_{S,i}$, feels full responsibility for the pair's earnings since their answer is implemented as the pair's final answer, and expects to disappoint the receiver by $m_{max} - m_{S,i}$.

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

$$\frac{m_{S,i} - m_{R,i}}{m_{max} - m_{S,i}} \geq w_{S,i} \quad (7)$$

¹¹ A sender's belief on how influential they are can be thought of an increasing function of ρ , such that $\frac{d\beta_{S,i}(\rho)}{d\rho} \geq 0$ and $\beta_{S,i}(1) = 1$. In the Dictator Treatment, $\rho = 1$, so $\beta_{S,i}(1) = 1$ for all subjects. This also indicates $m_{SR,i} = m_{S,i}$ because $x_{SR,i} = \beta_{S,i}(1)x_{S,i} + (1 - \beta_{S,i}(1))x_{R,i} = x_{S,i}$.

4.3 Effect of parameters on guess sending in advice and dictator treatments

To predict the effect of each mechanism outlined by the model, I examine the impact of varying the sender's self-confidence (affecting m_S and m_{SR}), their expectation of the receiver's performance (m_R), their sense of responsibility for their pair's earnings (Γ_S), their guilt intensity (w_S), and their belief in how much they influence the receiver (β_S).¹²

1. *Self-confidence* All else being equal, a sender with higher self-confidence is more likely to send their guess in both the Advice and Dictator Treatments compared to a sender with lower self-confidence. Higher confidence leads to an increase in both m_{SR} (increasing the left-hand side of Eq. 4) and m_S (increasing the left-hand side of Eq. 7).
2. *Expectation of the receiver's performance* A sender who expects higher performance from the receiver, and hence has a higher m_R , is less likely to send their guess in both treatments (since this leads to a decrease in the left-hand side of Eqs. 4 and 7).
3. *Responsibility for the pair's earnings* Having a higher sense of responsibility for the pair's earnings makes a sender less likely to send their guess in the Advice Treatment (by increasing the right-hand side of Eq. 4 through Γ_S) compared to a sender who feels less responsible for their pair. Since $\Gamma_S(1) = 1$ for all senders in the Dictator Treatment, guess-sending behavior would not differ between any two senders in the Dictator Treatment due to differences in their sense of responsibility.
4. *Guilt intensity* The model predicts that a sender with higher guilt intensity parameter is less likely to send their guess in both the Advice and Dictator treatments compared to a sender with lower guilt sensitivity (since a higher w_S increases the right-hand side of both Eqs. 4 and 7).
5. *Belief in their influence* A sender who believes that they are more influential over the receiver's answer is more likely to send their guess in the Advice Treatment compared to a sender with a lower belief (m_{SR} is increasing in β_S if the sender expects their answer to be more accurate than the receiver's and is decreasing in β_S otherwise. In the former case, a higher β_S increases the left-hand side of Eq. 4. In the latter case, the sender with the lower β_S would not be sending their guess to begin with, so an increase in β_S would not affect guess-sending behavior). Since $\beta_S = 1$ for all senders in the Dictator Treatment, guess-sending behavior would not differ between any two senders in the Dictator Treatment due to differences in their belief in their influence.

Table 2 summarizes the predicted effect of increasing each model parameter on guess-sending behavior in the Advice and the Dictator Treatments.

¹² In this section, I focus on differences between treatments at the subject level, so the parameters are aggregated at the subject level. For a discussion on how the parameters change at the question (or question difficulty) level, see Sect. 4.4.

4.4 Relationship between model parameters and question difficulty

In Subsection 4.3, I examine the effect of model parameters between treatments at the subject level. However, model parameters may vary by question, specifically based on question difficulty.¹³ In this subsection, I examine the effect of question difficulty on the model parameters.

Consider two questions, h and e , where question h is harder than question e . Firstly, subjects may have less self-confidence in their answers when the question is harder. This would lead to $m_{S,h} \leq m_{S,e}$ and $m_{SR,h} \leq m_{SR,e}$. At the same time, subjects may expect receivers to perform worse on harder questions compared to easier ones, resulting in $m_{R,h} \leq m_{R,e}$. Since $m_{R,i}$ only appears in the numerator, whereas $m_{S,i}$ and $m_{SR,i}$ appear in both the numerator and the denominator of the left-hand sides of equations 4 and 7, if the sender has a similar decrease in their expected earnings in harder questions when they send their guess compared to when they do not send their guess, the overall effect would result in less frequent guess-sending in harder questions compared to easier ones in both treatments.

Secondly, the difficulty of the question may affect subject's guilt through how responsible a subject feels for their pair's earnings or through their guilt intensity. Senders may feel more responsible when the problem is harder for the receiver to solve on their own. Conversely, senders may feel less responsible for the pair's earnings in harder questions because it is more challenging for them to know the answer as well, thus creating another excuse to shift the blame to the receiver. Similarly, subjects may feel more intense guilt in harder questions for letting down the receivers in questions they need more help with or less intense guilt because they go easier on themselves when the question is harder. Thus, the effect of difficulty through guilt can go in either direction. Any effect on responsibility, $\Gamma_{S,i}$, would have an impact only in the Advice Treatment, while any effect on guilt intensity, $w_{S,i}$, would have an impact in both treatments.

Lastly, senders may have a higher belief in influencing the receiver's answer in harder questions since it is more likely for receivers to have stronger opinions about the answer as the questions get easier. This would lead to $\beta_{S,h} \geq \beta_{S,e}$, resulting in senders being more likely to send their guess in harder questions compared to easier ones.

After presenting the main hypotheses of the experiment in Subsection 4.5 and experimental findings in Sects. 5 and 6, I will explore in Sect. 8 how the predictions of the guilt and responsibility model correspond with these findings and insights from the gender literature.

4.5 Hypothesis

The first question this paper addresses is whether there exists a gender gap in advice giving.

¹³ I would like to thank the anonymous reviewer #1 for pointing out the necessity of discussing the effect of question difficulty on model parameters and for the examples they provided.

Table 2 Predicted effect of increasing model parameters on guess sending behavior

Parameter to increase	Advice treatment	Dictator treatment
Self-confidence	More frequent	More frequent
Expectation of the receiver's performance	Less frequent	Less frequent
Responsibility for the pair's earnings	Less frequent	No change
Guilt intensity	Less frequent	Less frequent
Belief in their influence	More frequent	No change

Table summarizes the predicted effect of a sender sending their guess in the Advice and the Dictator Treatments in response to an increase in each parameter listed in the first column, according to Eqs. 4 and 7. Note that senders' responsibility for the pair's earnings and belief in their influence do not vary by design in the Dictator Treatment

Hypothesis 1 Female senders in the Advice Treatment are significantly less likely than male senders to send their guess to the receiver.

As outlined in Table 2, women having lower self-confidence, higher expectations of the receiver's performance, a greater sense of responsibility for the pair's earnings, more guilt intensity, or a lower belief in their influence on the receiver would all be in line with Hypothesis 1.

I experimentally vary the question difficulty with the intention of better understanding what drives advice giving behavior. I expect the subjects to know the correct answer without any meaningful effort in easy questions. Hence, easy questions can be seen as a sanity check to ensure that subjects were paying attention to the task and that they were sufficiently incentivized to send their advice. I expect the subjects to be almost sure that their answer is correct in easy questions. In medium questions, I expect subjects to count the correct number of red balls within the given time limit if they put effort into it. Therefore, a gender difference in advice sending in medium questions could indicate a gender difference in the effort levels between men and women in completing the task. Finally, the difficult questions were the ones in which I did not expect the subjects to be able to count the correct number of red balls within the given time limit. This is the question type in which I anticipate mechanisms other than payoff maximization to have an effect in advice sending behavior.

Hypothesis 2 Subjects have the lowest mean normalized error in easy questions and the highest mean normalized error in difficult questions.

If the questions classified as easy, medium, and difficult indeed correspond to these difficulties, I expect Hypothesis 2 to hold.

Hypothesis 3 Senders have the highest guess sending rate in easy questions and the lowest guess sending rate in difficult questions.

As discussed in Sect. 4.4, subjects might have lower self-confidence, higher sense of responsibility for the pair's earnings, and more intense guilt in harder questions,

making them more likely to send their guess for easy questions compared to difficult ones, in line with Hypothesis 3. On the other hand, subjects might have higher belief in influencing the receiver's answer, lower sense of responsibility for the pair's earnings, and less intense guilt in harder questions, working in the opposite direction and potentially leading to rejection of Hypothesis 3.

Hypothesis 4 The difference in the guess-sending rate between male and female senders is largest in difficult questions and smallest in easy questions in the Advice Treatment.

If women experience a greater decrease in their self-confidence, a greater increase (or smaller decrease) in feeling responsible, a greater increase (or smaller decrease) in their guilt intensity, or a smaller increase in their belief in influencing the receiver's guess in harder questions compared to easier ones relative to men, there would be a higher gender difference in difficult questions compared to easier ones in the Advice Treatment, in line with Hypothesis 4.

With the assumption that I find gender differences in guess sending in the Advice Treatment, I am then interested in understanding the factors contributing to the observed gender differences. One candidate I consider is that women may have lower self-confidence than men (as documented in Barber and Odean 2001, Niederle and Vesterlund 2007), making them less likely to give advice. I test this mechanism by controlling for self-confidence when examining advice giving behavior.

Hypothesis 5 Female senders have lower self-confidence compared to male senders.

All else being equal, Hypothesis 5 implies that women would be less likely to send their guess as advice.

Hypothesis 6 Gender difference in advice giving is not significant after controlling for self-confidence.

Hypothesis 6 can be rejected if the gender difference in self-confidence is not the only mechanism affecting the gender difference in advice giving.

Another possible mechanism is that women may dislike affecting others' choices more than men. For example, Ertac and Gurdal (2012) find that women do not like deciding on behalf of others in a risky task. If this is the underlying mechanism, making the advice enforceable would make guess sending even less desirable for women, potentially increasing the gender gap. Alternatively, men and women might view influencing others' choices similarly, but women might feel more responsible for the pair's earnings or have a lower belief that their advice will be listened to compared to men. In this case, making men and women equally responsible or equally influential would decrease the gender gap in advice-giving. To test this alternative explanation, I design an additional treatment in which I make advice enforceable.

Hypothesis 7 Both female and male senders are significantly less likely to send their guess in the Dictator Treatment compared to the Advice Treatment.

Senders' responsibility for the pair's earnings and influence on the receiver are higher in the Dictator Treatment compared to the Advice Treatment (since subjects are fully responsible for the pair's earnings and are fully influential on the receiver's answer if they send their guess in the Dictator Treatment). If the effect of the increase in the responsibility parameter outweighs the effect of the increase in the belief in influencing the receiver from the Advice to the Dictator Treatment, both men and women would be less likely to send their guess in the Dictator Treatment, in line with Hypothesis 7.

Hypothesis 8 The gender gap in guess-sending decreases when the advice becomes enforceable in the Dictator Treatment, with female senders experiencing a smaller decrease in their guess-sending compared to the Advice Treatment, in relation to male senders.

If women have a higher sense of responsibility than men in the Advice Treatment (which would decrease their incentive to send their guess in the Dictator Treatment less than men's) or a lower belief in how influential they are compared to men in the Advice Treatment (which would increase their incentive to send their guess in the Dictator Treatment more than men's), the gender gap would decrease in the Dictator Treatment, in line with Hypothesis 8.

5 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.¹⁴ 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.^{15,16} 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

¹⁴ 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.

¹⁵ 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.

¹⁶ See Table A.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.

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 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. I set certain cutoffs to define the difficulty of tasks (as can be seen in Table 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). In line with Hypothesis 2, subjects have the lowest mean normalized error in easy questions and the highest mean normalized error in difficult questions, indicating 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.¹⁷ 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 Sect. 6.4.

6 Experimental results

I begin by analyzing the senders' decision to send their guess in both treatments. Section 6.1 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 6.2 explores whether the gender gap persists in the Dictator Treatment, in which the senders' guess is enforced if they choose to send it. Section 6.3 compares the behavior across treatments. Finally, Sect. 6.4 examines the effect of advice on receivers' performance, both in the aggregate data and when the data is broken down by gender.

6.1 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. Figure 2 shows the percentages of all questions for which the senders sent their guesses, categorized by gender, and further broken down into easy, medium, and difficult questions. Pooling all questions together, men send advice in 83% of the questions, while women do

¹⁷ 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.

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. The gender difference in advice sending is not statistically significant for easy ($p = 0.112$) 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 A.1 plots the cumulative distribution functions (CDFs) of advice sending percentages for male and female senders in all questions, as well as questions broken down by difficulty in the Advice Treatment. The first order stochastic dominance test using Somers' D statistic (Newson et al., 2001) indicates that there is a significant dominance relationship between the distributions of male and female senders at the 95% confidence level, but this significance is observed only for difficult questions. Based on Somers' D statistic, a randomly chosen male sender is 27% more likely to send advice for a difficult question than a female sender ($p = 0.018$), supporting the first order stochastic dominance relationship. No significant dominance

relationship is found in the case of easy, medium, or pooled questions at the 95% confidence level.¹⁸

Additional evidence for gender differences in guess sending can be found in Table 3. This table presents the results of the Probit regressions that investigate the relationship between guess sending in the Advice Treatment and gender, as well as interactions between gender and each difficulty level (the excluded category is female senders and difficult questions). Standard errors are clustered at the individual level. The bottom part of the table reports p -values associated with the F -tests, testing the statistical significance of gender differences in guess sending for each difficulty level separately, as well as testing for their joint significance.

The regression results support the previous finding from non-parametric tests that men are significantly more likely than women to send their guess as advice in difficult questions ($p = 0.016$ for the F -test testing $H_0 : \beta_{Male} = 0$). There is no gender difference in guess sending in medium questions ($p = 0.649$ for the F -test testing $H_0 : \beta_{Male} + \beta_{Male \times Medium} = 0$), also similar to the non-parametric findings. In easy questions, Probit regressions suggest that women are significantly more likely to send their guess in the Advice Treatment compared to men ($p = 0.021$ for the F -test testing $H_0 : \beta_{Male} + \beta_{Male \times Easy} = 0$),

whereas no significance was documented in the non-parametric tests.¹⁹ Furthermore, gender difference is significant when the null hypotheses for the gender difference in each difficulty level is tested jointly (p -value associated with the joint F -test $p = 0.004$). The results are robust to controlling for individual performance

¹⁸ Somer's D statistic (indicating the likelihood of a randomly chosen male sender sending their guess relative to a female sender in the Advice Treatment) and the associated p -values for each difficulty level are as follows: all questions: 20%, $p = 0.074$, difficult questions: 27%, $p = 0.018$, medium questions: 5%, $p = 0.604$, easy questions: -8%, $p = 0.168$.

¹⁹ The results are qualitatively similar when I run separate regressions for each difficulty level. I report the results of these regressions as a robustness check in Sect. 7.

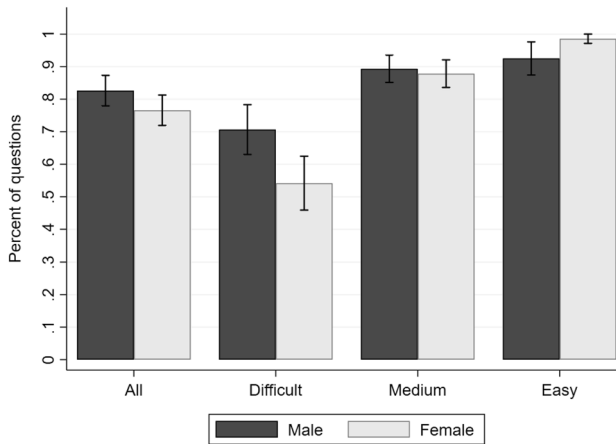


Fig. 2 Percent of Questions for which the Senders Send Their Guess, Advice Treatment. *Notes:* Figure illustrates percentages of all, difficult, medium, and easy questions for which the senders send their guess, broken down by gender. The p -values for the differences of percentages between men and women are $p = 0.068$ for all questions (without a difficulty breakdown), $p = 0.014$ for difficult, $p = 0.601$ for medium, and $p = 0.112$ for easy questions

(measured by the sender's normalized error on a question), self-confidence (measured by their self-perceived rank of guess accuracy), demographics, and risk preferences, as shown in the second column of Table 3.

Since male and female 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 in guess sending in difficult questions is not actual performance differences. Table A.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 3 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.²⁰ 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.

²⁰ 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).

Comparing the results in this section to the hypotheses outlined in Sect. 3, I find support for Hypothesis 1 and show that question difficulty is an important factor for gender differences in advice giving. Consistent with Hypothesis 4, the gender difference in guess-sending is most pronounced and significant in difficult questions, with women being less likely than men to send their guess. The difference is smaller and not significant in medium questions, and it even reverses in easy questions, with women being more likely than men to send their guess. Additionally, I find support for Hypothesis 5, women having lower self-confidence than men. However, despite self-confidence being a significant factor in guess-sending, the gender gap persists after controlling for self-confidence, thereby rejecting Hypothesis 6.

6.2 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), the gender gap in guess-sending disappears when senders can enforce their guess. Figure 3 plots the percentages of all questions for which the senders sent their guesses, categorized by gender, and further broken down into easy, medium, and difficult questions. Pooling all difficulty levels together, men send their guess in 72% of the questions, while women do so in 69% of the questions. The difference is not statistically significant ($p = 0.210$). 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).

Figure A.2 plots the cumulative distribution functions (CDFs) of advice sending percentages for male and female senders in all questions, as well as questions broken down by difficulty in the Dictator Treatment. The first order stochastic dominance test using Somers' D statistic (Newson et al., 2001) indicates that there is no significant dominance relationship between the distributions of male and female senders at the 95% confidence level for any difficulty level in the Dictator Treatment.²¹

Table 4 provides further evidence that there is no gender difference in guess sending in any difficulty

level in the Dictator Treatment. This table presents the results of the Probit regressions that investigate the relationship between guess sending in the Dictator Treatment and gender, as well as the interaction between gender and each difficulty level (the excluded category is female senders and difficult questions). Standard errors are clustered at the individual level. The bottom part of the table reports p -values

²¹ Somer's D statistic (indicating the likelihood of a randomly chosen male sender sending their guess relative to a female sender in the Dictator Treatment) and the associated p -values for each difficulty level are as follows: all questions: 14%, $p = 0.224$, difficult questions: 8%, $p = 0.484$, medium questions: 12%, $p = 0.272$, easy questions: -4%, $p = 0.478$.

Table 3 Probit regressions relating sender's guess-sending to gender in advice treatment

Guess sent	(1)	(2)
Male	0.437** (0.016)	0.405** (0.036)
Medium	1.054*** (0.000)	1.035*** (0.000)
Easy	2.081*** (0.000)	2.061*** (0.000)
Male x medium	−0.353** (0.030)	−0.330* (0.056)
Male x easy	−1.167*** (0.001)	−1.195*** (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	0.108 (0.384)	0.430 (0.323)
<i>p-values associated with F-tests, testing gender difference in guess sending:</i>		
Joint	0.004	0.003
Difficult	0.016	0.036
Medium	0.649	0.708
Easy	0.021	0.013
N	2,773	2,773

Top half of the table Reports coefficients and *p*-values from the Probit regression. 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), *Easy* and *Medium* (dummy variables indicating whether question difficulty was easy and medium, respectively), gender and difficulty interactions (the

Table 3 (continued)

excluded category is female & difficult), *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 and 4). Errors are clustered at the individual level

Bottom half of the table Reports *p*-values associated with the *F*-tests, testing the statistical significance of gender differences in guess sending for each difficulty level separately, as well as testing for their joint significance. *Difficult* tests $H_0 : \beta_{Male} = 0$, *Medium* tests $H_0 : \beta_{Male} + \beta_{Male \times Medium} = 0$, *Easy* tests $H_0 : \beta_{Male} + \beta_{Male \times Easy} = 0$, and *Joint* tests all three hypotheses jointly

p-values are reported in parentheses * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

associated with the *F*-tests, testing the statistical significance of gender differences in guess sending for each difficulty level separately, as well as testing for their joint significance.

The regression results support the previous finding from non-parametric tests that there is no gender difference in guess sending in any difficulty level in the Dictator Treatment (*p*-values associated with *F*-tests testing the gender difference in difficult, medium, and easy questions are 0.410, 0.279, and 0.677, respectively), in line with the findings in the non-parametric tests.²² Furthermore, gender difference is not significant when the null hypotheses for the gender difference in each difficulty level is tested jointly (*p*-value associated with the joint *F*-test is 0.526). The results are robust to controlling for individual performance (measured by the sender's normalized error on a question), self-confidence (measured by their self-perceived rank of guess accuracy), demographics, and risk preferences, as shown in the second column of Table 4.

The results in this section show that, the gender difference in guess-sending disappears when the advice becomes enforceable in the Dictator Treatment, in line with Hypothesis 8.

6.3 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. Table 5 illustrates percentage of guesses sent by treatment and gender

²² The results are qualitatively similar when I run separate regressions for each difficulty level. I report the results of these regressions as a robustness check in Sect. 7.

when questions are pooled at the difficulty level. Senders send their guess in 80% of the questions in the Advice Treatment, whereas they do so in 71% of the questions in the Dictator Treatment (difference is statistically significant, $p < 0.001$). Breaking down the analysis by gender shows that, in line with Hypothesis 7, both female and male senders are significantly less likely to send their guess in the Dictator Treatment compared to the Advice Treatment ($p = 0.016$ for female and $p < 0.001$ for male).

Breaking down questions by difficulty: Figure A.3 shows that senders send their guess in 96% and 98% of the easy questions, 89% and 87% of the medium questions, and 62% and 41% of the difficult questions in the Advice and Dictator Treatments, respectively. The difference in guess sending across treatments is significant in difficult questions ($p < 0.001$), marginally significant in medium questions ($p = 0.053$), and not significant in easy questions ($p = 0.524$). Figure A.4 plots the cumulative distribution functions (CDFs) of guess-sending frequency across treatments. The CDF in the Advice Treatment first order stochastically dominates the CDF in the Dictator Treatment only for difficult questions (and in pooled questions when there is not a difficulty breakdown) at the 95% significance level. Based on Somers' D statistic, a randomly chosen sender in the Dictator Treatment is 38% less likely to send their guess in a difficult question than a sender in the Advice Treatment ($p < 0.001$).²³

Guess-sending behavior by difficulty levels in each treatment shows that, in line with Hypothesis 3, senders are significantly less likely to send their guesses as the questions become harder. The frequency of guess-sending differs significantly between medium and difficult questions, as well as between easy and medium questions, in both the Advice and the Dictator Treatments ($p < 0.001$). Furthermore, the change in guess-sending by difficulty level varies between treatments. The average decrease in guess-sending from medium to difficult questions is 26% compared to 46% ($p < 0.001$), and from easy to medium questions, it is 7% compared to 11% ($p = 0.017$) in the Advice and the Dictator Treatments.

Figure 4 illustrates the percentages of guesses sent by the senders for all, difficult, medium, and easy questions, categorized by treatment and gender. Analyzing guess-sending by difficulty level suggests that difficult questions are the main drivers of differences in guess sending across treatments. The gender difference in guess sending between the Advice Treatment and the Dictator Treatment is smaller and not statistically significant for easy or medium questions, for both genders, in comparison to difficult questions and overall questions without a difficulty breakdown.

For the difficult questions, both male and female senders send their guess at significantly lower rates in the Dictator Treatment compared to the Advice

²³ Somer's D statistic (indicating the likelihood of a randomly chosen sender in the Dictator Treatment sending their guess relative to a sender in the Advice Treatment) and the associated p -values for each difficulty level are as follows: all questions: -33% , $p < 0.001$; difficult questions: -38% , $p < 0.001$; medium questions: -14% , $p = 0.058$; easy questions: 2% , $p = 0.594$.

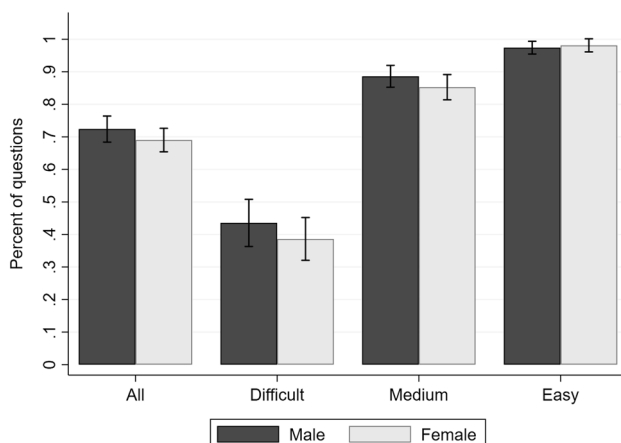


Fig. 3 Percent of Questions For Which Senders Send Their Guess, Dictator Treatment. *Notes:* Figure illustrates percentages of all, difficult, medium, and easy questions for which the senders send their guess, broken down by gender. The p -values for the differences of percentages between men and women are $p = 0.210$ for all questions (without a difficulty breakdown), $p = 0.478$ for difficult, $p = 0.265$ for medium, and $p = 0.565$ for easy questions

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. 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.

6.4 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 77% of the questions overall, and 94%, 86%, and 60% of the rounds for easy, medium, and difficult questions, respectively. In this section, I present the results for all difficulty levels for completeness, but it's important to note that the frequency of advice in easy and medium questions is quite high. Therefore, the results of this section are most relevant for difficult questions.

In the Advice Treatment, the receivers' average normalized error is 11.5% in rounds without advice, while it is 4.2% in rounds with advice. Receivers perform significantly better in rounds for which they receive advice ($p < 0.001$). When questions are broken down by difficulty level, the only significant difference in receivers' performance with and without advice is in difficult questions. In difficult questions, the receivers' average normalized error in rounds with and without advice is 8.7% and 13.7%, respectively ($p < 0.001$). In medium questions, average normalized error

Table 4 Probit Regressions
Relating Sender's Guess-
Sending to Gender in Dictator
Treatment

Guess Sent	(1)	(2)
Male	0.124 (0.410)	0.022 (0.876)
Medium	1.346*** (0.000)	1.356*** (0.000)
Easy	2.387*** (0.000)	2.349*** (0.000)
Male x medium	0.034 (0.828)	0.029 (0.864)
Male x easy	-0.261 (0.455)	-0.300 (0.453)
Period		0.005 (0.195)
Error		-0.019*** (0.000)
Risk averse		0.220 (0.158)
High education		-0.051 (0.665)
Employed		-0.079 (0.539)
Age		-0.007 (0.196)
Rank guess: 2		-0.661*** (0.000)
Rank guess: 3		-1.018*** (0.000)
Rank guess: 4		-0.649 (0.125)
Constant	-0.291*** (0.005)	0.546* (0.059)
<i>p-values associated with F-tests, testing gender difference in guess sending:</i>		
Joint	0.526	0.719
Difficult	0.410	0.876
Medium	0.279	0.740
Easy	0.677	0.478
N	2,762	2,762

Top half of the table Reports coefficients and *p*-values from the Probit regression. 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), *Easy* and *Medium* (dummy variables indicating whether question difficulty was easy and medium, respectively), gender and difficulty interactions (the

Table 4 (continued)

excluded category is female & difficult), *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 and 4). Errors are clustered at the individual level

Bottom half of the table Reports *p*-values associated with the *F*-tests, testing the statistical significance of gender differences in guess sending for each difficulty level separately, as well as testing for their joint significance. *Difficult* tests $H_0 : \beta_{Male} = 0$, *Medium* tests $H_0 : \beta_{Male} + \beta_{Male \times Medium} = 0$, *Easy* tests $H_0 : \beta_{Male} + \beta_{Male \times Easy} = 0$, and *Joint* tests all three hypotheses jointly

p-values are reported in parentheses; * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 5 Percentage of guesses sent by treatment and gender

Treatment	Pooled	Female	Male	Gender <i>p</i> -value
Advice	0.80	0.77	0.83	0.068
Dictator	0.71	0.69	0.72	0.210
Treatment <i>p</i> -value	<0.001	0.016	<0.001	

There are 112 observations (56 male, 56 female) in both treatments. The column "Gender *p*-value" and the row "Treatment *p*-value" report the *p*-values associated with the gender and treatment differences in guess sending being significantly different from 0, respectively, using a Wilcoxon Rank Sum Test

with and without advice is 2.9% and 6.2% ($p = 0.145$), while in easy questions, it is 0.8% and 0.3% ($p = 1.000$).

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 documented performance difference in rounds with and without advice in difficult questions could be driven by omitted variable bias rather than advice improving decision making. To test this explanation, I examine how the receivers' performance in difficult questions 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 A.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.²⁴ In line with the advice literature, (e.g., Schotter, 2003; Cooper & 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.²⁵ The average normalized error of men (women) is 4.9% (6.4%) for all questions without a difficulty breakdown, 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, and 0.045 for all questions). 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 arise only in rounds for which advice is sent. Figure A.5 breaks down the average normalized errors by gender, difficulty, and whether the receiver received advice and rejects this hypothesis. Pooling all difficulty levels together, the results show that the gender difference in receiver performance actually arises from rounds without advice, rather than rounds with advice. The difference in average normalized error between men and women is not statistically significant (3.9% for men, 4.5% for women, $p = 0.402$) in rounds in which the receiver receives advice. On the contrary, the difference is significant in rounds without advice (9.0% for men, 14.0% for women, $p = 0.006$). Breaking down questions by difficulty, a similar pattern arises, specifically in difficult questions.²⁶ The findings suggest that

²⁴ Since the decision to send advice differs between the Advice treatment and the Dictator treatment, endogeneity concern cannot be fully ruled out.

²⁵ 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.

²⁶ In difficult questions, the difference in average normalized error between men and women is not statistically significant (8.4% for men, 9.0% for women, $p = 0.450$) in rounds with advice, and the difference is significant in rounds without advice (11.6% for men, 15.8% for women, $p = 0.010$). In medium questions, men outperform women in both rounds with advice (2.6% for men, 3.2% for women, $p = 0.019$) and without advice (2.9% for men, 9.2% for women, $p = 0.043$), but the gender difference in performance is larger in rounds without advice. In easy questions, the gender difference in average normalized error between men and women is marginally significant (0.5% for men, 1.2% for women, $p = 0.051$) in rounds with advice and not significant (0.1% for men, 0.8% for women, $p = 0.759$) in rounds without advice. The results on medium and easy questions in rounds without advice should be interpreted with caution, since most receivers have received advice in medium and easy questions, so the sample sizes are small for those cases.

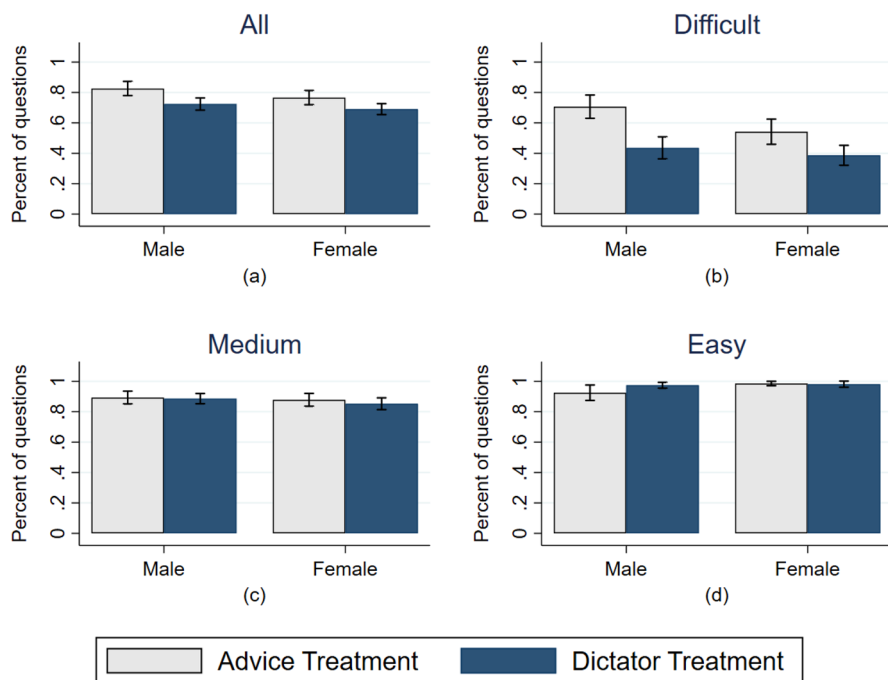


Fig. 4 Percentage of Questions For Which Senders Send Their Guess Across Treatments. *Notes:* Figure illustrates the percentage of questions for which senders sent their guesses, broken down by difficulty, treatment, and gender. The p -values for the difference in percentages between the Advice and Dictator Treatments for male and female senders are as follows: $p < 0.001$ (male) and $p = 0.016$ (female) for all questions, depicted in Panel (a), $p < 0.001$ (male) and $p = 0.024$ (female) for difficult questions, depicted in Panel (b), $p = 0.266$ (male) and $p = 0.106$ (female) for medium questions, depicted in Panel (c), and $p = 0.392$ (male) and $p = 0.904$ (female) for easy questions, depicted in Panel (d)

the gender gap in performance, especially in difficult questions, is driven by rounds without advice, falsifying the hypothesis that men incorporate advice better in these questions. 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 for this differential effect of advice on men and women.

To investigate who gives better advice, I compare the average normalized error of male and female senders, conditional on sending their guesses. I do not find any gender difference in the quality of advice across treatments. Pooling all questions together, the average normalized error of male and female senders are 6.7% and 4.7%, respectively, in the Advice Treatment, while it is 3.0% and 4.4% in the Dictator Treatment. The F-statistic testing the joint significance of gender differences across all difficulty levels indicates that men and women do not have significantly different accuracy in their sent guesses ($p = 0.380$ in the Advice Treatment, $p = 0.106$

in the Dictator Treatment).²⁷ One interesting finding is that, conditional on sending their guesses, men send more accurate guesses in the Dictator Treatment than in the Advice Treatment (F-statistic testing the joint significance of performance differences across all difficulty levels for men yields $p = 0.005$); while women do not increase the quality of their guesses when the advice becomes enforceable (the difference in average normalized errors across treatments $p = 0.335$).²⁸ 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 the enforceability of advice.

Finally, I examine whether being matched with a male or a female sender leads to better answers by the receivers. I find no significant difference in the average normalized error based on the sender's gender in either treatment. In the Advice Treatment, the average normalized error of receivers with male and female senders is 5.5% and 5.8%, respectively, while in the Dictator Treatment, it is 6.2% and 6.1%. The F-statistic testing the joint significance of sender's gender differences across all difficulty levels indicates that having a male or female sender does not lead to significantly different performance of receivers ($p = 0.890$ in the Advice Treatment, $p = 0.986$ in the Dictator Treatment). Breaking down questions by difficulty, the effect of having a male or female sender remains insignificant for all difficulty levels ($p > 0.1$). 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.²⁹

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

²⁷ Breaking down questions by difficulty, there is no gender difference in average normalized errors in difficult questions (Advice Treatment: 10.4% and 9.5% for men and women, $p = 0.531$; Dictator Treatment: 7.4% and 8.9% for men and women, $p = 0.346$), as well as in medium questions (Advice Treatment: 6.5% and 4.1% for men and women, $p = 0.111$; Dictator Treatment: 2.4% and 3.5% for men and women, $p = 0.221$). The only performance difference across genders arises for easy questions in the Dictator Treatment, with men having a lower average normalized error (Advice Treatment: 1.3% and 0.5% for men and women, $p = 0.301$; Dictator Treatment: 0.0% and 2.4% for men and women, $p = 0.028$).

²⁸ When questions are broken down by difficulty, the result remains similar, especially in difficult and medium questions: men have significantly better performance in rounds that they send their guess in the Dictator Treatment compared to the Advice Treatment ($p = 0.026$ for difficult, $p = 0.002$ for medium, $p = 0.053$ for easy questions), whereas women's performance remains statistically similar across treatments ($p = 0.709$ for difficult, $p = 0.629$ for medium, $p = 0.095$ for easy questions).

²⁹ The results are similar if I compare payoffs instead of average normalized errors for all questions in the Advice Treatment and for difficult and medium questions in the Dictator Treatment ($p > 0.1$). Having a male sender does significantly increase the pair's payoffs in easy questions and all questions without a difficulty breakdown ($p < 0.001$) in the Dictator Treatment. The difference in payoffs in easy questions in the Dictator Treatment is in line with male senders' higher performance in easy questions in the Dictator Treatment in rounds that they send advice. For other difficulty levels, it is possible that given the percentage point gender difference in advice sending, the effect of receiving advice 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.

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' behavior (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.

7 Robustness analysis

7.1 Varying the cutoff for classifying a question as difficult

In this subsection and the following one, I conduct robustness analyses to ensure that the main result—women being less likely to send advice than men in difficult questions—is not artificially influenced by the specific cutoffs chosen to classify the difficulty of questions.

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 $\bar{\delta}$ as the cutoff such that questions with $\delta > \bar{\delta}$ are classified as “difficult” questions. In the main analysis, I used $\bar{\delta} = 30$ as the cutoff to classify questions as “difficult”. In this section, I vary the $\bar{\delta}$ 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”.³⁰

Figure 5 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 $\bar{\delta}$ used as a cutoff to classify whether a question is “difficult”. The case where $\bar{\delta} = 0$ 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 all cutoff values $\bar{\delta}$, 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 A.7 shows how many questions are classified as “difficult” for each possible value of $\bar{\delta}$, along with the percentage of advice sent by men and women, the gender difference in percentage of advice sent, and the

³⁰ The reason for varying $\bar{\delta}$ up to 48 is that for higher values of $\bar{\delta}$, 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.

p -value associated with the gender difference in advice sending when question difficulty is determined by the corresponding $\bar{\delta}$. The gender gap in advice sending becomes significant at the 5% level when all but the easiest 4 questions are classified as “difficult” (when $\bar{\delta} = 7$, 21 out of 25 questions are classified as “difficult”). The difference remains positive and mostly increasing for higher values of $\bar{\delta}$. 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 $\bar{\delta} \geq 46$, 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.

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 Sect. 7.1), as an alternative to grouping questions in three categorical difficulty levels, which was the analysis conducted in Sect. 6.1.

Table A.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.

7.3 Separate regressions for each difficulty level

In this subsection, I present regression results for difficult, medium, and easy questions separately as an alternative to analyzing advice giving by interacting the gender with difficulty levels and applying F -test to test gender differences in guess sending at each difficulty level.

Table A.4 presents the results of the Probit regressions relating guess sending to gender separately for each difficulty level in the Advice Treatment. These findings confirm the earlier results from Sect. 6.1, which utilized a single regression with all difficulty levels, incorporating gender and difficulty interactions. In difficult questions, men are significantly more likely to send their guess compared

to women ($p = 0.016$). There is no gender difference in guess sending in medium questions ($p = 0.649$). Lastly, in easy questions, men are significantly less likely to send their guess compared to women ($p = 0.021$). The results are similar after controlling for sender's performance, self-confidence, and risk aversion as well as period, education, employment, and age.

Table A.5 reports the same analysis for the Dictator Treatment. These findings further confirm the earlier results obtained using a single regression with gender and difficulty interactions in Sect. 6.2. There is no gender difference in guess sending at any difficulty level (p -values are 0.410, 0.279, and 0.677 for difficult, medium, and easy questions, respectively). The results remain similar after controlling for sender's performance, self-confidence, risk aversion, period, education, employment, and age.

Hence, the results in Sects. 6.1 and 6.2 are robust to examining behavior in question difficulties in isolation as an alternative to pooling all questions together and interacting gender with difficulty levels.

8 Discussion

This paper documents that women are less likely than men to send advice, and that question difficulty is an important factor contributing to this gender gap. In fact, female senders are more likely than men to give advice on the easiest questions, whereas they become significantly less likely to do so as the questions get more difficult. The gender difference cannot be explained by performance, as both genders are equally successful when they do not receive any external information in the task. Moreover, the two additional mechanisms considered in this experiment, self-confidence and disliking to decide on behalf of others, are not sufficient to explain the gender difference in advice-giving. I find that although female subjects have lower self-confidence than males, and self-confidence affects the decision to give advice, the gender difference still persists after controlling for it. Guess-sending in an environment where the guesses are enforceable and subjects fully decide on behalf of others closes the gender gap, rather than increasing it.

The model of guilt and responsibility outlined in Sect. 4 lays out several candidate mechanisms for the gender gap documented in this paper. Comparing the model predictions in Table 2 with the gender literature and the experimental results, some mechanisms come forward. Firstly, women are shown to have lower self-confidence than men (in this paper and also in others, e.g. Barber and Odean, 2001; Niederle and Vesterlund, 2007), so gender differences in self-confidence can contribute to women being less likely to send advice. However, self-confidence alone cannot be the only mechanism leading to the gender difference, as it would result in an even stronger gender difference in the Dictator Treatment. Secondly, Manian and Sheth (2021) show that subjects do not expect women's advice to be followed as much as men's. If women believe that their advice will not be followed as much as men's, this would make them less likely to send their guess in the Advice Treatment, in line with the experimental results. Third, several studies find that men exhibit higher

guilt aversion than women (Nihonsugi et al., 2022; Di Bartolomeo et al., 2022).³¹ Hence, the gender differences in guilt intensity documented in the literature would predict an opposite effect to the one found in this paper. Lastly, even though there are no studies specifically focusing on gender differences in blame-shifting, 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, men would be more likely to send their guess in the Advice Treatment, where there is room to shift the blame, but no gender difference would arise in the Dictator Treatment, since subjects are fully responsible for their pair's earnings in this treatment by design.

In terms of the effect of question difficulty on the willingness to send advice, Subsection 4.4 shows that question difficulty may also have opposing effects on the willingness to give advice. The experimental findings reveal that subjects in both treatments become significantly less willing to send their guess as the questions get harder. This indicates that the effect of having lower self-confidence likely dominates the effect of expecting the receiver to have worse performance as the questions get harder. The drop in guess-sending from easier to harder questions is larger in the Dictator Treatment compared to the Advice Treatment, suggesting that difficulty may influence either feeling responsibility for the pair's earnings or belief in influencing the receiver's answer, which are the two mechanisms asymmetrically impacting the Advice and the Dictator Treatments. Moreover, the gender gap in guess sending reverses from easy questions to difficult questions, but only in the Advice Treatment. One possible reason for this might be that women experience a larger drop in self-confidence from easy to difficult questions. However, this mechanism by itself cannot explain the finding because it would require women to have higher self-confidence than men in easy questions, and the effect would be similar in the Dictator Treatment. Differences in how the sense of responsibility or belief in influencing changes from easy to difficult questions for men and women can potentially explain the findings. Women might feel less responsible for the pair's earnings in easy questions but experience a larger increase in their sense of responsibility from easy to difficult questions. Or, women might believe their advice is more likely to be followed in easy questions but experience a steeper drop in that belief from easy to difficult questions. Both of these mechanisms would result in the gender difference reversing in the Advice Treatment from easy to difficult questions, but would not have the same effect in the Dictator Treatment since the sense of responsibility and belief in influence do not vary by question difficulty in this treatment.³²

As indicated by the discussion above, the experimental design of this paper highlights several candidates, but it cannot determine the mechanism leading to the gender difference in advice-giving. This is because several features considered in the model change simultaneously when moving from the Advice to the Dictator

³¹ 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 .

³² I thank the anonymous reviewer #1 for the very helpful examples they provided on why the gender difference possibly varies by question difficulty, but only in the Advice Treatment.

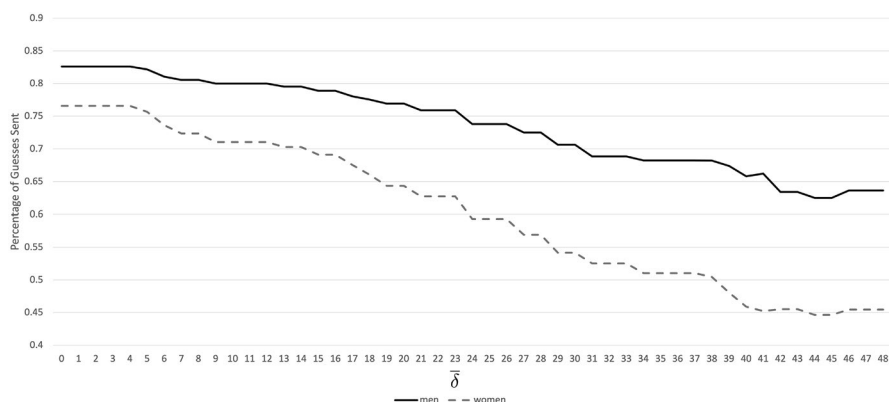


Fig. 5 Percentage of Advice Sending in Difficult Questions by Gender for Different Values of $\bar{\delta}$. *Notes:* Figure illustrates the percentage of questions for which senders send their guess in difficult questions in the Advice Treatment, broken down by gender. The x-axis varies the $\bar{\delta}$ cutoff used to classify a question as “difficult”. At $\bar{\delta} = 0$, all questions are classified as difficult. At $\bar{\delta} = 48$, only one question (with 49 red balls and 51 blue balls) is classified as difficult

Treatment. Beliefs about improving decision-making through advice, beliefs about advice-following, and the extent of shifting the blame for (sub-optimal) group earnings are features that play a role in the Advice Treatment, but not in the Dictator Treatment. Beliefs about own relative performance, responsibility aversion for the pair’s earnings, and guilt aversion for (sub-optimal) group outcome are features present in both treatments, but have a heavier weight in the Dictator Treatment.³³ The model introduced in Sect. 4 aims to serve as a guide for future research, rather than determining the exact mechanism behind the gender differences.

9 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, the gender gap in guess-sending disappears. Both men and women send their guess significantly less in the Dictator Treatment, but the decrease is greater for men, diminishing the gender gap.

³³ I thank the anonymous reviewer #2 for their thoughtful suggestions about the differences across the treatments.

This paper also supports findings in the advice literature indicating that the presence of advice enhances the performance of receivers (e.g., Schotter 2003, Cooper and Kagel 2016). Additionally, it documents 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 among 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 females in the Advice Treatment. This performance gap is driven by rounds in which the receivers did not receive advice, suggesting that the performance difference is not due to men being better at following advice, but rather presence of an advisor having a differential effect on how men and women perform in rounds where they do not receive advice.

The focus of this paper was to investigate the existence of a gender gap in advice-giving and its interaction with task difficulty. The natural next step is to identify the exact mechanism leading to these gender differences in advice-giving. I have introduced a model of guilt and responsibility, outlining several candidate mechanisms that could contribute to the documented gender gap. However, the current experimental design does not allow for the identification of the specific mechanism responsible for the gender gap. Designing an experiment to change one feature at a time, while eliciting subjects' levels of guilt, responsibility for their pair's earnings, self-confidence at the question level, beliefs about the receiver's performance, and beliefs about their influence on the receiver's answer, can help pinpoint the exact mechanism. This remains a topic for future research.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s10683-024-09846-w>.

References

- Babcock, L., Recalde, M. P., Vesterlund, L., & Weingart, L. (2017). Gender differences in accepting and receiving requests for tasks with low promotability. *American Economic Review*, 107(3), 714–47.
- Barber, B. M., & Odean, T. (2001). Boys will be boys: Gender, overconfidence, and common stock investment. *The Quarterly Journal of Economics*, 116(1), 261–292.
- Bartling, B., & Fischbacher, U. (2012). Shifting the blame: On delegation and responsibility. *The Review of Economic Studies*, 79(1), 67–87.
- Battigalli, P., & Dufwenberg, M. (2007). Guilt in games. *American Economic Review*, 97(2), 170–176.
- Battigalli, P., Charness, G., & Dufwenberg, M. (2013). Deception: The role of guilt. *Journal of Economic Behavior & Organization*, 93, 227–232.
- Beyer, S. (1990). Gender differences in the accuracy of self-evaluations of performance. *Journal of Personality and Social Psychology*, 59(5), 960.
- Bohren, J. A., Imas, A., & Rosenberg, M. (2019). The dynamics of discrimination: Theory and evidence. *American economic review*, 109(10), 3395–3436.
- Born, A., Ranehill, E., & Sandberg, A. (2020). Gender and willingness to lead: Does the gender composition of teams matter? *The Review of Economics and Statistics*, 104, 1–46.
- Brandts, J., & Rott, C. (2021). Advice from women and men and selection into competition. *Journal of Economic Psychology*, 82, 102333.
- Brandts, J., Groenert, V., & Rott, C. (2015). The impact of advice on women's and men's selection into competition. *Management Science*, 61(5), 1018–1035.

- Catalyst, P. (2020). Women in s & p 500 companies. Available at SSRN 3239814
- Celen, B., Kariv, S., & Schotter, A. (2010). An experimental test of advice and social learning. *Management Science*, 56(10), 1687–1701.
- Chakraborty, P., Serra, D., et al. (2021). Gender and leadership in organizations: Promotions, demotions and angry workers. In: *Working Papers* 20210104–001
- Charness, G., & Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, 74(6), 1579–1601.
- Chaudhuri, A., Schotter, A., & Sopher, B. (2009). Talking ourselves to efficiency: Coordination in inter-generational minimum effort games with private, almost common and common knowledge of advice. *The Economic Journal*, 119(534), 91–122.
- Chen, D. L., Schonger, M., & Wickens, C. (2016). otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9, 88–97.
- Chen, J., & Houser, D. (2019). When are women willing to lead? the effect of team gender composition and gendered tasks. *The Leadership Quarterly*, 30(6), 101340.
- Coffman, K., Flikkema, C. B., & Shurchkov, O. (2021). Gender stereotypes in deliberation and team decisions. *Games and Economic Behavior*, 129, 329–349.
- Coffman, K. B. (2014). Evidence on self-stereotyping and the contribution of ideas. *The Quarterly Journal of Economics*, 129(4), 1625–1660.
- Coffman, K. B., Exley, C. L., & Niederle, M. (2021). The role of beliefs in driving gender discrimination. *Management Science*, 67(6), 3551–3569.
- Cooper, D. J., & Kagel, J. H. (2016). A failure to communicate: An experimental investigation of the effects of advice on strategic play. *European Economic Review*, 82, 24–45.
- Di Bartolomeo, G., Martin, D., Papa, S., Laura, R., et al. (2022). *Guilt aversion: Eve versus adam*. In *Guilt Aversion: Eve versus Adam*.
- Ding, T., & Schotter, A. (2017). Matching and chatting: An experimental study of the impact of network communication on school-matching mechanisms. *Games and Economic Behavior*, 103, 94–115.
- Ding, T., & Schotter, A. (2019). Learning and mechanism design: An experimental test of school matching mechanisms with intergenerational advice. *The Economic Journal*, 129(623), 2779–2804.
- Eckel, C. C., & Grossman, P. J. (2008). Men, women and risk aversion: Experimental evidence. *Handbook of experimental economics results*, 1, 1061–1073.
- Erat, S. (2013). Avoiding lying: The case of delegated deception. *Journal of Economic Behavior & Organization*, 93, 273–278.
- Ertac, S., & Gurdal, M. Y. (2012). Deciding to decide: Gender, leadership and risk-taking in groups. *Journal of Economic Behavior & Organization*, 83(1), 24–30.
- Eskreis-Winkler, L., Milkman, K. L., Gromet, D. M., & Duckworth, A. L. (2019). 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.
- Exley, C. L., & Kessler, J. B. (2019). *The gender gap in self-promotion*. National Bureau of Economic Research: Technical report.
- Gneezy, U., & Potters, J. (1997). An experiment on risk taking and evaluation periods. *The Quarterly Journal of Economics*, 112(2), 631–645.
- Gneezy, U., Saccardo, S., Serra-Garcia, M., & van Veldhuizen, R. (2020). Bribing the self. *Games and Economic Behavior*, 120, 311–324.
- Heikensten, E., Isaksson, S. (2019). Simon says: Examining gender differences in advice seeking and influence in the lab. Available at SSRN 3273186
- Hinnosaar, M. (2019). Gender inequality in new media: Evidence from wikipedia. *Journal of Economic Behavior & Organization*, 163, 262–276.
- Iyengar, R., & Schotter, A. (2008). Learning under supervision: an experimental study. *Experimental Economics*, 11(2), 154–173.
- Lazear, E. P., & Rosen, S. (1990). Male-female wage differentials in job ladders. *Journal of Labor Economics*, 8(1), S106–S123.
- Manian, S., & Sheth, K. (2021). Follow my lead: Assertive cheap talk and the gender gap. *Management Science*, 67(11), 6880–6896.
- Newson, R., et al. (2001). somersd-confidence intervals for nonparametric statistics and their differences. *Stata Technical Bulletin*, 10 (55)
- Niederle, M., & Vesterlund, L. (2007). Do women shy away from competition? do men compete too much? *The Quarterly Journal of Economics*, 122(3), 1067–1101.
- Nihonsugi, T., Tanaka, T., & Haruno, M. (2022). Gender differences in guilt aversion in korea and the united kingdom. *Scientific Reports*, 12(1), 8187.

- Nyarko, Y., Schotter, A., & Sopher, B. (2006). On the informational content of advice: A theoretical and experimental study. *Economic Theory*, 29(2), 433–452.
- Peeters, R., & Vorsatz, M. (2021). Simple guilt and cooperation. *Journal of Economic Psychology*, 82, 102347.
- Rothenhäusler, D., Schweizer, N., & Szech, N. (2018). Guilt in voting and public good games. *European Economic Review*, 101, 664–681.
- Schotter, A. (2003). Decision making with naive advice. *American Economic Review*, 93(2), 196–201.
- Zhou, H., & Fishbach, A. (2016). 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.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.