Gender Differences in the Cost of Corrections in Group Work

Yuki Takahashi*

Click here for the latest version

April 11, 2022

Abstract

Having a good relationship with colleagues is integral in group work, potentially leading to successful collaborations. However, there are occasions when people have to correct their colleagues. I study whether people dislike collaborating with someone who corrects them and more so when that person is a woman. Using a quasi-laboratory experiment, I find that people, including those with high productivity, are less willing to collaborate with a person who has corrected them even if the correction improves group performance. In addition, I find suggestive evidence that men respond more negatively to women's corrections, which is not driven by their beliefs about the difference in women's and men's abilities. These findings suggest that there is a behavioral bias in group work that distorts the optimal selection of talents and penalizes those who correct others' mistakes, and the distortion may be stronger when women correct men.

JEL codes: J16, M54, D91, C92
Keywords: correction, collaboration, group work, gender differences, quasi-laboratory experiment

*Department of Economics, University of Bologna. Email: yuki.takahashi2@unibo.it. I am grateful to Maria Bigoni, Siri Isaksson, and Bertil Tungodden, whose feedback was essential for this project, and to the experiment participants for their participation and cooperation. This paper also benefited from helpful comments by Laura Anderlucci, Seda Ertaç, Valeria Ferraro, Boon Han Koh, Annalisa Loviglio, Valeria Maggian, Natalia Montinari, Vincenzo Scrutinio, participants at the CSQIEP Job Market Seminar, Stanford Institute for Theoretical Economics conference, Webinar in Gender and Family Economics, seminars at Ca' Foscari University, Norwegian School of Economics, Osaka University, the University of Bologna, and many other people. Tommaso Batistoni, Philipp Chapkovski, Christian König genannt Kersting, and oTree help & discussion group kindly answered my questions about oTree programming; in particular, my puzzle code was heavily based on Christian's code. Francesca Cassanelli, Natalia Montinari, and Ludovica Spinola helped me to write experimental instructions in Italian. Michela Boldrini and Boon Han Koh conducted the quasi-laboratory experiments ahead of me and kindly answered my questions about the implementations. Lorenzo Golinelli provided excellent technical and administrative assistance. This study was pre-registered with the OSF registry (https://osf.io/tgyc5) and approved by the IRB at the University of Bologna on November 3, 2020 (ref. no. 262643).

1 Introduction

Having a good relationship with colleagues is vital in group work, and one key benefit is having successful collaboration opportunities. Since people do most work in groups both in the industry (Lazear and Shaw 2007) and in the academia (Jones 2021; Wuchty, Jones, and Uzzi 2007), successful collaboration opportunities are essential in their career.

However, on some occasions, one has to correct their colleague; one instance of such occasions is when a colleague at a business meeting suggests something against the group's interests (e.g., an investment plan that likely has a negative net present value) or interests of some group members (e.g., setting a meeting time later in the afternoon when nursery schools are already closed, which burdens people with small children). Another instance of such occasion is when a colleague presents their paper at an academic seminar which has some potential flaws in the identification strategy or the experimental design.

Such corrections may not affect one's relationship with their colleagues if those colleagues are rational agents,[1] but in reality, corrections may negatively affect one's relationship with their colleagues because people avoid information that conflicts with their own beliefs (Abelson 1986; Golman, Hagmann, and Loewenstein 2017). In particular, colleagues may dislike women's corrections more because people often use double standards for women and men (Egan, Matvos, and Seru 2021; Sarsons 2019; Sinclair and Kunda 2000).[2]

This paper studies whether people dislike collaborating with someone who corrects them and more so when that person is a woman. Answering this question using secondary data poses two challenges. First, group formation is not random and group corrections are endogenous. Second, different corrections are not necessarily comparable to each other.

To overcome these challenges, I design a quasi-laboratory experiment, a hybrid of physical laboratory and online experiments, where group formation is randomized and define corrections such that researchers can track its quality mathematically. Specifically, participants are allocated to a group of eight and solve one joint task with each group member one by one. Each time participants finish the task, they state whether they would like to collaborate with the group member with whom they have just solved the task for the same task in the next stage, which is the main source of earnings. This gives a strong incentive for participants to select as good a collaborator as possible. The order of the group members with whom participants solve the task is randomized. As a joint task, I use Isaksson (2018)'s number-sliding puzzle, which allows me to calculate an objective measure of each participant's contribution to the joint task as well as to classify each move as good (move the puzzle closer to the solution) or bad (move the puzzle further away from the solution). I define a correction as reversing a group member's move, which is comparable across different participants and can be classified as either good or bad.

1. Information avoidance is sometimes rational; see (Golman, Hagmann, and Loewenstein 2017).

2. Evidence suggests that men undervalue women when they criticize them (Sinclair and Kunda 2000) and that people punish women more harshly when they make mistakes (Sarsons 2019) and commit misconduct (Egan, Matvos, and Seru 2021).

I find that people correctly understand the notion of good and bad moves; that is, the higher your contribution is to solving the puzzle, the more likely it is that you will be asked to join a team. This is in line with what one would expect and validates my experimental design.

Nonetheless, after controlling for the contribution, people are less willing to collaborate with a person who has corrected their moves, even if the corrections move the puzzle closer to the solution. Although it may not be so costly to correct colleagues if only low productivity people respond negatively to corrections because collaborations with those people will likely result in low quality outcomes, high productivity people also respond negatively to corrections. High productivity people's negative response also suggests that the negative response is likely to be irrational: they should be able to better identify good and bad corrections.

Although only suggestive, I also find evidence that men respond more negatively to women's good corrections: that is, men may dislike women correct their mistakes. This finding is unlikely to be due to people's beliefs about the differences in women's and men's abilities in the puzzle: women and men contribute equally well to the puzzle, and neither women nor men underestimate women's contribution.

Taken together, these findings suggest that there is a behavioral bias that distorts the optimal selection of talents and penalizes those who correct others' mistakes, and men may exhibit stronger bias when women correct them.

This paper's contribution is twofold. First, it contributes to the growing literature on the effect of workplace climate and productivity. For example, Alan, Corekcioglu, and Sutter (2021) find that a better workplace climate increases worker satisfaction and the degree of mutual reciprocation while reducing toxic competition and worker turnover, and argue that improved manager-worker relationships are the likely mechanism. Collaborating this evidence, Edmans (2011) finds that firms with high employee satisfaction exhibit higher stock prices, and Guiso, Sapienza, and Zingales (2015) find that a firm performs better when its workers perceive their managers as trustworthy and ethical. Also, workplace climate can differentially affect women and men: Dupas et al. (2021) find that female economists receive more patronizing and hostile questions during seminars. I show that interpersonal frictions can distort group efficiency in a controlled laboratory environment where I directly relate corrections and decisions to collaborate, and the friction may have a stronger effect on women.

Second, it contributes to the literature on gender differences in the contribution of ideas. For example, Coffman (2014) finds that women are less likely to contribute their ideas to the group in a male task due to self-stereotyping, and Gallus and Heikensten (2019) find that debiasing their self-stereotyping by giving an award for their high ability increases women's contribution: they put women's idea further ahead of men without involving open correction of their group member. However, on some occasions, one has to contribute their ideas openly, for example, at business meetings and academic seminars. In such cases, group members' response plays an important role in the effectiveness of the intervention suggested by Gallus and Heikensten (2019), and indeed there is evidence that people respond to women's ideas less favorably: Coffman, Flikkema, and Shurchkov

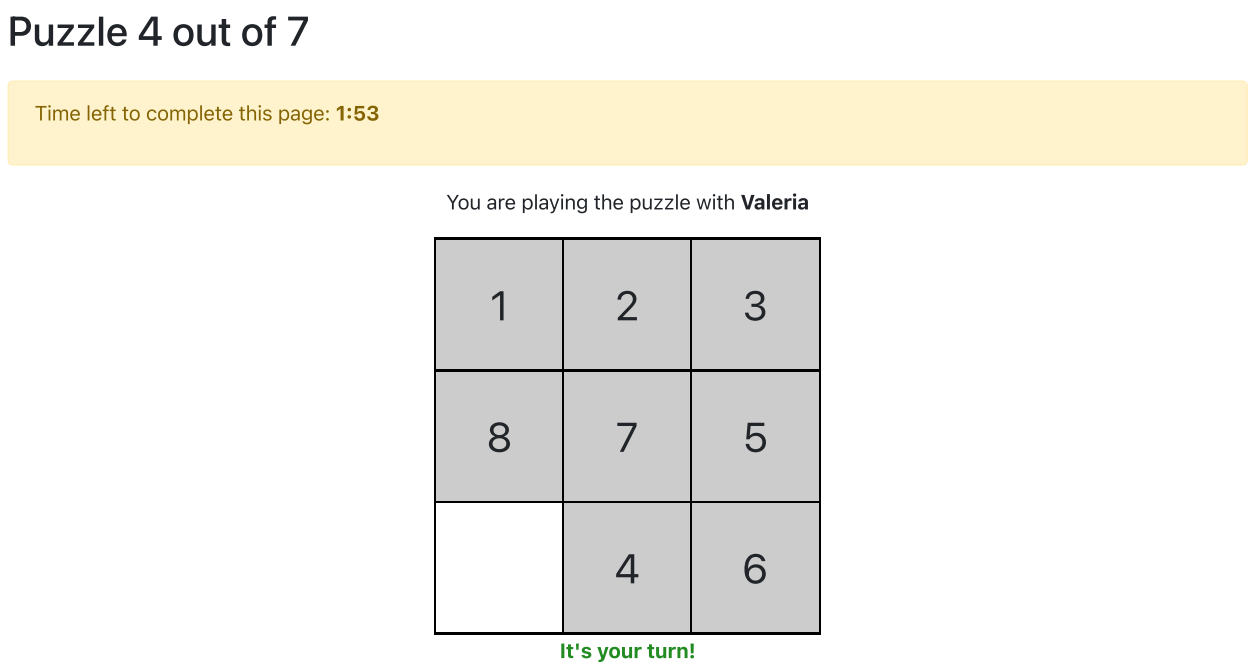
(2021) find that group members are less likely to choose women's answers as a group answer in male-typed questions. Corroborating this, Guo and Recalde (2020) find that group members correct women's ideas more often than men's ideas. I introduce correction in the contribution of ideas and examine its cost on women and groups so that we can design more effective interventions.

The remainder of the paper proceeds as follows. Section 2 describes the experimental design, procedure, and implementation. Section 3 describes the data. Section 4 presents a simple theoretical framework. Section 5 presents evidence that people are less willing to collaborate with a person who has corrected their moves, even if the corrections move the puzzle closer to the solution. Section 6 presents evidence that even high productivity people respond negatively to corrections. Section 7 presents suggestive evidence that men respond more negatively to women's good corrections. Section 8 presents the robustness of the results. Section 9 concludes.

2 Experiment

Introducing a quasi-laboratory format I run the experiment in a quasi-laboratory format where we experimenters connect us to the participants via Zoom throughout the experiment (but turn off participants' camera and microphone except at the beginning of the experiment) and conduct it as we usually do in a physical laboratory, but participants participate remotely using their computers. Appendix A discusses the advantages and drawbacks of the quasi-laboratory format relative to physical laboratory and standard online experiments.

Figure 1: Puzzle screen



Notes: This shows a sample puzzle screen where a participant is matched with another participant called Valeria at the 4th round of the puzzle and making their move. All the texts are in Italian in the experiment.

Group task As the group task, I use Isaksson (2018)'s puzzle, a sliding puzzle with eight numbered tiles, which should be placed in numerical order within a 3x3 frame (see Figure 1 for an example). To achieve this goal, participants play in pairs, alternating their moves.[3] This puzzle has nice mathematical properties: I can define the puzzle difficulty and classify a given move as either good or bad by the Breadth-First Search algorithm.[4] From the number of good and bad moves one makes, I can calculate individual contributions to the group task; I measure it by net good moves, the number of good moves minus the number of bad moves an individual makes in a given puzzle.

I can also determine the quality of corrections of different participants objectively and comparably.[5] Further, the puzzle-solving captures an essential characteristic of teamwork in which two or more people work towards the same goal (Isaksson 2018), but the quality of each move and correction is only partially observable to participants (but fully observable to the experimenter).

At each stage of the puzzle, there is only one good strategy which is to make a good move and one bad strategy which is to make a bad move.[6] There can be more than one good and bad move, but different good/bad moves are equal. There is no path dependence either: the history of the puzzle moves does not matter.

At the beginning of each part, participants must answer a set of comprehension questions to make sure they understand the instructions.[7]

2.1 Design and procedure

Registration

Upon receiving an invitation email to the experiment, participants register for a session they want to participate in and upload their ID documents as well as a signed consent form.[8]

Pre-experiment

On the day and the time of the session they have registered for, participants enter the Zoom waiting room.[9] They receive a link to the virtual room for the experiment and enter their first name, last name, and their email they have used in the registration. They also draw a virtual coin numbered from 1 to 40 without replacement.

3. Each participant has to make a move in their turn; they cannot pass.

4. The difficulty is defined as the number of moves away from the solution, a good move is defined as a move that reduces the number of moves away from the solution, and a bad move is defined as a move that increases the number of moves away from the solution.

5. Indeed, some corrections happen early in the puzzle and the other later in the puzzle. Thus, what I capture in the analysis is the average effect of a correction.

6. This is conditional on that both players are trying to solve the puzzle; I show in section 8 that the results are robust to exclusion of puzzles where either player might not be trying to solve the puzzle.

7. I do not tell participants that they can correct others to reduce experimenter demand effects.

8. I recruit a few more participants than I would need for a given session in case some participants would not show up to the session.

9. Zoom link is sent with an invitation email; I check that they have indeed registered for a given session before admitting them to the Zoom meeting room.

Then I admit participants to the Zoom meeting room one by one and rename them by the first name they have just entered. This information is necessary to match up their earnings in this experiment and their payment information stored in the laboratory database, so participants have a strong incentive to provide their true name and email address. If there is more than one participant with the same first name, I add a number after their first name (e.g., Giovanni2).

After admitting all the participants to the Zoom meeting room, I do roll call, a way to reveal participants' gender to other participants without making gender salient (Bordalo et al. 2019; Coffman, Flikkema, and Shurchkov 2021). Specifically, I take attendance by calling each participant's first name one by one and ask her or him to respond via microphone. This process ensures other participants that the called participant's first name corresponds to their gender. If there are more participants than I would need for the session (I need 16 participants), I draw random numbers from 1 to 40 and ask those who drew the coins with the same number to leave.[10] Those who leave the session receive the 2 show-up fee. Figure 2 shows a Zoom screen participants would see during the roll call (the person whose camera is on is the experimenter; participants would see this screen throughout the experiment, but the experimenter's camera may be turned off).

I then read out the instructions about the rules of the experiment and take questions on Zoom. Once participants start the main part, they can communicate with the experimenter only via Zoom's private chat.

Part 1: Individual practice stage

Participants work on the puzzle individually with an incentive (0.2 for each puzzle they solve). They can solve as many puzzles as possible with increasing difficulty (maximum 15 puzzles) in 4 minutes. This part familiarizes them with the puzzle and provides us with a measure of their ability given by the number of puzzles they solve. After the 4 minutes are over, they receive information on how many puzzles they have solved.

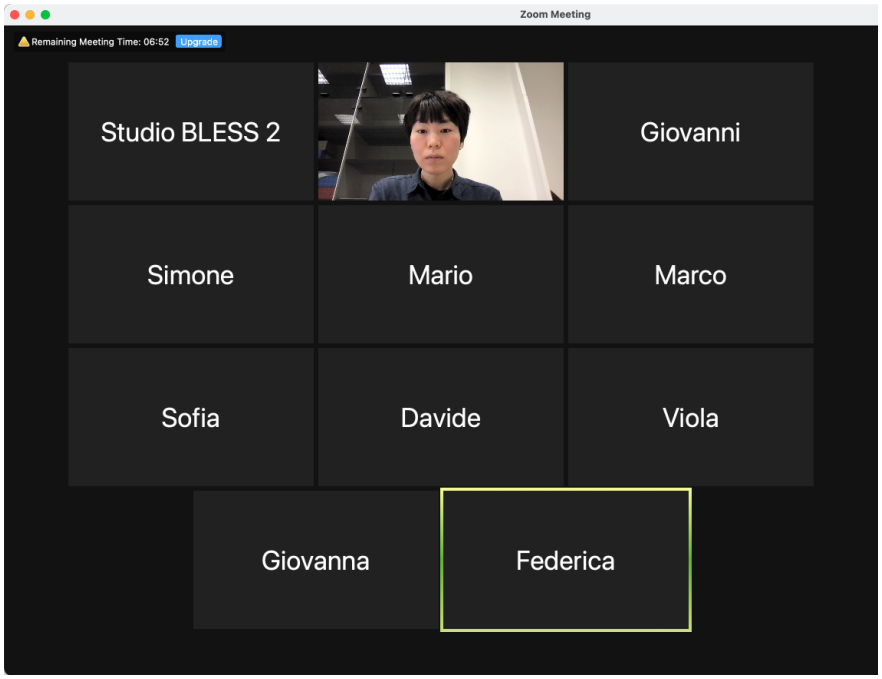
Part 2: Collaborator selection stage

Part 2 contains seven rounds, and participants learn the rules of part 3 before starting part 2. This part is based on Fisman et al. (2006, 2008)'s speed dating experiments and proceeds as follows: first, participants are allocated to a group of 8 based on their ability similarity as measured in part 1. This is done to reduce ability difference among participants, and participants do not know this grouping criterion.

Second, participants are paired with another randomly chosen participant in the same group and solve one puzzle together by alternating their moves. The participant who makes the first move is drawn at random and both participants know this first-mover selection criterion. If they cannot

10. I draw with replacement a number from 1 to 40 using Google's random number generator (https://www.google.com/search?q=random+number). If no participant has a coin with the drawn number, I draw next number until the number of participants is 16. I share my computer screen so that participants see the numbers are actually drawn randomly.

Figure 2: Zoom screen



Notes: This figure shows a Zoom screen participants would see during the roll call. The experimenter's camera is on during the roll call. Participants would see this screen throughout the experiment but the experimenter's camera may be turned off.

solve the puzzle within 2 minutes, they finish the puzzle without solving it. Participants are allowed to reverse the paired participant's move.[11] Reversing the partner's move is what I call correction in this paper. Each participant's contribution in a given puzzle is measured by net good moves. Figure 1 shows a sample puzzle screen where a participant is paired with another participant called Valeria and making their move.[12] The paired participant's first name is displayed on the computer screen throughout the puzzle and when participants select their collaborator to subtly inform the paired participant's gender.

Once they finish the puzzle, participants state whether they would like to collaborate with the same participant in part 3 (yes/no). At the end of the first round, new pairs are formed, with a perfect stranger matching procedure, so that every participant is paired with each of the other seven members of their group once and only once. In each round, participants solve another puzzle in a pair, then state whether they would like to collaborate with the same participant in part 3. The sequence of puzzles is the same for all pairs in all sessions. The puzzle difficulty is kept the same across the seven rounds. The minimum number of moves to solve the puzzles is set to 8 based on the pilot.

11. Solving the puzzle itself is not incentivized, and thus participants who do not want to collaborate with the paired participant or fear to receive a bad response may not reverse that participant's move even if they think the move is wrong. However, since I am interested in the effect of correction on collaborator selection, participants' *intention* to correct that does not end up as an actual correction does not confound the analysis.

12. All the texts are in Italian in the experiment.

At the end of part 3, participants are paired according to the following algorithm:

1. For every participant, call it i, I count the number of matches; that is, the number of other participants in the group who were willing to be paired with i and with whom i is willing to collaborate in part 3.
2. I randomly choose one participant.
3. If the chosen participant has only one match, I pair them and let them work together in part 3.
4. If the chosen participant has more than one match, I randomly choose one of the matches.
5. I exclude two participants that have been paired and repeat (1)-(3) until no feasible match is left.
6. If some participants are still left unpaired, I pair them up randomly.

Part 3: Group work stage

The paired participants work together on the puzzles by alternating their moves for 12 minutes and earn 1 for each puzzle solved. Which participant makes the first move is randomized at each puzzle, and this is told to both participants as in part 2. They can solve as many puzzles as possible with increasing difficulty (maximum 20 puzzles).

Post-experiment

Each participant answers a short questionnaire which consists of (i) the six hostile and benevolent sexism questions used in Stoddard, Karpowitz, and Preece (2020) with US college students and (ii) their basic demographic information and what they have thought about the experiment.[13] The answer to their demographic information is used to know participants' characteristics as well as casually check whether they have anticipated that the experiment is about gender, for which I do not find any evidence.

After participants answer all the questions, I tell them their earnings and let them leave the virtual room and Zoom. They receive their earnings via PayPal.

2.2 Implementation

The experiment was programmed with oTree (Chen, Schonger, and Wickens 2016) and conducted in Italian during November-December 2020. I recruited 464 participants (244 female and 220 male) registered on the Bologna Laboratory for Experiments in Social Science's ORSEE (Greiner 2015) who (i) were students, (ii) were born in Italy, and (iii) had not participated in gender-related experiments before (as far as I could check).[14] The first two conditions were to reduce noise coming from

13. I was planning to use a gender bias measure constructed from the hostile and benevolent sexism questions to show those with higher gender bias respond more negatively to women's corrections. However, people do not respond more negatively to women's corrections and that I could not have enough variation in this gender bias measure, so decided not to report it. See Appendix B for the results.

14. The laboratory prohibits deception, so no participant has participated in an experiment with deception.

differences in socio-demographic backgrounds and race or/and ethnicity that may be inferred from participants' first name or/and voice, and the last condition was to reduce experimenter demand effects. The number of participants was determined by a power simulation in the pre-analysis plan to achieve 80% power.[15] The experiment is pre-registered with the OSF.[16]

I ran 29 sessions with 16 participants each. The average duration of a session was 70 minutes. The average total payment per participant was 11.55 with the maximum 25 and the minimum 2, all including the 2 show-up fee. Table 1 describes participants' characteristics. The table shows that while female participants are slightly younger (1.41 years) and less gender biased (0.12). In addition, female participants are more likely to major in humanities and male participants are more likely to major in natural sciences and engineering, a tendency observed in most OECD countries (see, for example, Carrell, Page, and West 2010).[17] Also, most female and male participants are either bachelor or master students (97% of female and 94% of male) and only few are PhD students.

Table 1: Participants' characteristics

	Female (N=244)			Male (N=220)			Difference (Female – Male)	
	Mean	SD	Median	Mean	SD	Median	Mean	P-value
Age	24.45	3.13	24	25.87	4.33	25	-1.41	0.00
Gender bias	0.17	0.16	0.12	0.29	0.19	0.29	-0.12	0.00
Region of origin:								
North	0.32			0.36			-0.04	0.37
Center	0.23			0.24			-0.01	0.77
South	0.45			0.40			0.06	0.23
Abroad	0.00			0.00			0.00	0.32
Major:								
Humanities	0.45			0.22			0.23	0.00
Social sciences	0.24			0.27			-0.03	0.52
Natural sciences	0.12			0.20			-0.08	0.02
Engineering	0.05			0.23			-0.17	0.00
Medicine	0.13			0.08			0.05	0.08
Program:								
Bachelor	0.34			0.26			0.08	0.06
Master	0.63			0.68			-0.05	0.26
Doctor	0.03			0.06			-0.03	0.11

Notes: This table describes participants' characteristics. P-values of the difference between female and male participants are calculated with heteroskedasticity-robust standard errors.

15. This number includes 16 participants from a pilot session run before the pre-registration where the experimental instructions were slightly different. The results are robust to exclusion of these 16 participants.

16. The pre-registration documents are available at the OSF registry: https://osf.io/tgyc5.

17. Individual fixed effects in the analysis control for one's major. However, I do not run heterogeneity analysis by major because major choice is endogenous to one's gender.

3 Data description

I use part 2 data in the analysis as part 2 is where we can observe collaborator selection decisions. I aggregate the move-level data at each puzzle so that we can associate behaviors in the puzzle to the collaborator selection decisions. As shown in Figure 3, both mixed gender and single-gender groups perform equally well (panel A), about 71% of the puzzles are solved within a minimum number of moves (panel B, the minimum number of moves is 8), and corrections happen across the moves (panel C).

Table 2: Own and partners' puzzle behaviors and puzzle outcomes

	Female (N=1708)		Male (N=1540)		Difference (Female – Male)		
	Mean	SD	Mean	SD	Mean	SE	P-value
Panel A: Own behaviors							
Contribution	2.98	2.93	3.14	2.64	-0.16	0.10	0.11
# puzzles solved in part 1	8.36	2.41	8.80	2.34	-0.44	0.22	0.05
Any correction	0.15	0.36	0.16	0.36	0.00	0.01	0.85
Good correction	0.12	0.33	0.12	0.33	0.00	0.01	0.90
Bad correction	0.06	0.23	0.05	0.22	0.00	0.01	0.70
Panel B: Partner's behaviors							
Contribution	3.04	2.73	3.07	2.87	-0.03	0.10	0.77
# puzzles solved in part 1	8.58	2.35	8.57	2.43	0.01	0.16	0.93
Any correction	0.16	0.37	0.15	0.36	0.01	0.01	0.51
Good correction	0.13	0.33	0.12	0.32	0.01	0.01	0.44
Bad correction	0.06	0.23	0.05	0.22	0.01	0.01	0.44
Panel C: Puzzle outcomes							
Willing to collaborate (yes=1, no=0)	0.72	0.45	0.71	0.45	0.01	0.02	0.49
Time spent (second)	43.74	36.15	42.99	35.76	0.74	1.28	0.56
Total moves	11.18	7.46	11.21	7.70	-0.03	0.28	0.92
Puzzle solved	0.85	0.36	0.86	0.35	-0.01	0.01	0.43
Consecutive correction	0.04	0.20	0.04	0.21	0.00	0.01	0.81

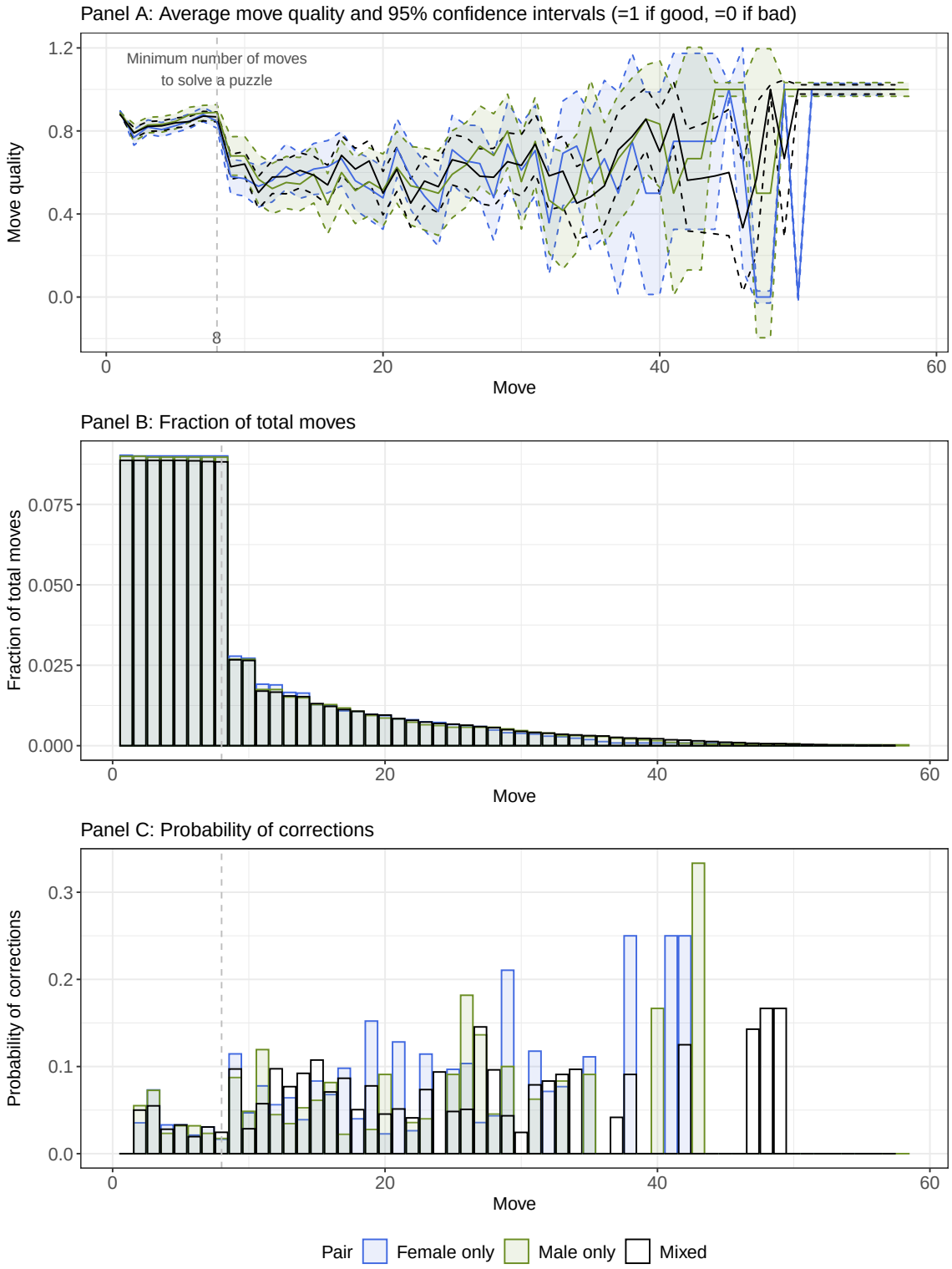
Notes: This table describes own (panel A) and partner's puzzle behaviors (panel B) and puzzle outcomes (panel C). P-values of the difference between female and male participants are calculated with standard errors clustered at the individual level. Contribution is defined as one's net good moves in a given puzzle (the number of good moves minus the number of bad moves).

Table 2 describes own (panel A) and partner's puzzle behaviors (panel B) and puzzle outcomes (panel C). Panel A shows that there are no gender differences in puzzle-solving ability: both contribution in part 2 and the number of puzzles solved in part 1, the difference between female and male participants are statistically insignificant at 5% and quantitatively insignificant.[18][19] This is consistent with Isaksson (2018), who also finds no gender difference in contribution or number of

18. The number of puzzles solved in part 1 is marginally significant but quantitatively insignificant.

19. The correlation coefficient between contribution and number of puzzles solved in part 1 is 0.1059 and the p-value is below 0.001 (with standard errors clustered at individual level).

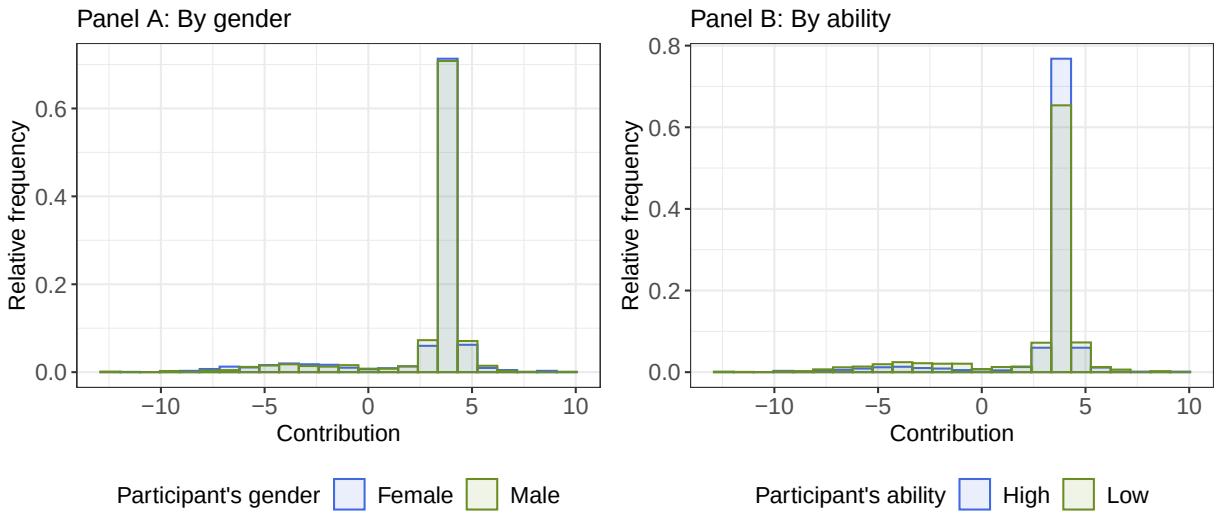
Figure 3: Move quality, fraction of total moves, and probability of corrections



Notes: The average move quality along with 95% confidence intervals (panel A), the fraction of total moves in each move (panel B), and the probability of corrections in each move (panel C), separately for female only (gray), male only (white), and mixed gender pairs (blue). The confidence interval of panel A is 95% confidence intervals of βs from the following OLS regression: $MoveQuality_{ijt} = \beta_1 + \sum_{k=2}^{58} \beta_k \mathbb{1}[t_{ij} = k] + \epsilon_{ijt}$, where t_{ij} is the pair i-j's move round and $\mathbb{1}$ is an indicator variable. $MoveQuality_{ijt}$ takes a value of 1 if a move of a pair i-j in tth move is good and 0 if bad. I add an estimate of β_1 to estimates of β_2-β_{58} to make the figure easier to look at. Standard errors are clustered at the pair level.

puzzles solved alone using the same puzzle, suggesting that any gender difference I would find is unlikely to come from their ability difference. Panel A also shows that there are no gender differences in propensity to correct partners, suggesting any gender differences I would find are not coming from either gender corrects more than the other gender.

Figure 4: Distribution of contribution



Notes: This figure shows the distribution of individual contribution by gender (panel A) and ability (panel B) and shows that most participants contributed to the same degree. Panel A further shows no gender difference in contribution, and panel B further shows that among high-ability people, higher fraction contributes to the puzzles to the same degree. Contribution is defined as one's net good moves in a given puzzle (the number of good moves minus the number of bad moves).

Panel A of Figure 4 presents the distribution of contribution by participants gender to further elaborate panel A of Table 2 that women and men are equally good at puzzle solving: in about 70% of the puzzles, participants' contribution is 4 (total good moves minus total bad moves), and women's and men's distributions almost overlap.

Panel B shows that puzzle-solving ability as well as propensity to make corrections (both of a mistake and of a right move) of partners paired with female and male participants is the same, suggesting random pairing was successful and that any gender differences I would find are not coming from partners of either gender correct more often. Participants are corrected by their partner in 15-16% of the total puzzles, of which 12-13% are good corrections, and 5-6% are bad corrections.[20]

Panel C shows that participants state they want to collaborate with the partner 71-72% of the time. Participants spend on average 43-44 seconds for each puzzle (the maximum time a pair can spend is 120 seconds), and take 11 moves. 85-86% of the puzzles are solved and participants and the partner correct each other's move consecutively in 4% of the puzzles.[21] There is no gender difference

20. The percentage of good corrections and bad corrections do not sum up to the percentage of any correction means there are puzzles where both good and bad corrections occurred. The results are robust to exclusion of these overlapping puzzles, as shown in Figures 6, 7, and 8.

21. Indeed, in puzzles where consecutive correction happens, probability of selecting a paired participant as collaborator drops from 78.0% to 26.8%.

in any of these outcomes, suggesting any gender differences cannot be attributed to the imbalance in these outcomes.[22]

4 Theoretical framework

I present a simple theoretical framework to provide a benchmark for rational agent's behaviors.

I consider a fully rational participant i who maximizes their expected payoff by selecting their collaborator j from a set of i's potential collaborators $J \equiv \{1, 2, 3, 4, 5, 6, 7\}$. The payoff is increasing in i's belief about j's ability. I assume i can partially observe j's move quality, so i's belief about j's ability is increasing in j's ability perceived by i.

Thus, if i would select with whom to play in part 3, i would face the following problem:

$$\max_{j \in J} E_{\mu_j}[\pi(\mu_j(\tilde{a}_j, c_j, f_j))|\theta_i, \omega_i], \quad \partial\pi/\partial\mu_j > 0,\ \partial\mu_j/\partial\tilde{a}_j > 0 \tag{1}$$

where each term is defined as follows:

- μ_j: i's belief about j's ability
- $\tilde{a}_j$: j's ability perceived by i
- c_j: j's correction (=1 if j corrected i, =0 if j did not correct i)
- f_j: j's gender (=1 if female, =0 if male)
- θ_i: i's belief about their ability relative to other participants (>0 if higher, =0 if same, <0 if lower)
- ω_i: j's belief about women's ability relative to men (>0 if higher, =0 if same, <0 if lower)

If i can fully observe j's move quality and i is fully rational, then j's correction, c_j, and gender, f_j, do not convey any information about j's ability and is irrelevant for i's decision making. However, since i can only partially observe j's move quality, j's correction and gender convey information about j's ability even if i is fully rational.[23]

First, keeping j's ability perceived by i fixed, the information j's correction conveys depends on θ_i. If i believes they are good at the puzzle, they would consider a correction as a signal of low ability because i believes their move is correct. On the other hand, if i believes their ability is low, then they would consider a correction as a signal of high ability. If i believes their ability is the same as j's, then a correction would not convey any information.

However, since i can partially observe j's move quality, i considers a good correction as less negative/more positive signal than a bad corrections regardless of θ_i. Thus, we have the following proposition:

22. Note that time spent to solve a puzzle is endogenous to correction and not a good control. For example. if one corrects a mistake, then it takes fewer time to solve the puzzle. If one corrects a right move, on the other hand, then it takes more time to solve the puzzle.

23. I nonparametrically control for j's gender, but I also examine the effect of interaction term between j's correction and j's gender.

Proposition 1. *A fully rational participant responds more negatively to a bad correction than to a good correction regardless of their belief about their ability relative to other participants. That is:*

$$\partial\mu_j/\partial c_j|_{c_j \text{ is a bad correction}} < \partial\mu_j/\partial c_j|_{c_j \text{ is a good correction}} \forall\theta_i \tag{2}$$

Also, the more the i understands the puzzle, the more they can observe j's move quality, hence corrections, regardless of θ_i. Thus, we have the following proposition:

Proposition 2. *A fully rational participant with higher puzzle solving ability respond less negatively to good corrections and more negatively to bad corrections than another fully rational participant with lower puzzle solving ability. This is true regardless of their belief about their ability relative to other participants. That is:*

$$\begin{aligned} \partial\mu_j/\partial c_j|_{i\text{'s ability is high} \wedge c_j \text{ is a good correction}} &> \partial\mu_j/\partial c_j|_{i\text{'s ability is low} \wedge c_j \text{ is a good correction}} \forall\theta_i \\ \partial\mu_j/\partial c_j|_{i\text{'s ability is high} \wedge c_j \text{ is a bad correction}} &< \partial\mu_j/\partial c_j|_{i\text{'s ability is low} \wedge c_j \text{ is a bad correction}} \forall\theta_i \end{aligned} \tag{3}$$

Similar to response to corrections, if i believes women is better at the puzzle, they would consider a correction from a woman as a signal of high ability relative to men's correction. On the other hand, if i believes women is worse at the puzzle, then they would consider a correction from a woman as a signal of low ability relative to men's correction. If i believes women and men are equally good at the puzzle, then a correction from a woman or man is irrelevant. Thus, we have the following proposition:

Proposition 3. *A fully rational participant's response to women's corrections relative to men's correction depends on their belief about women's ability relative to men's. This is true regardless of their belief about their ability relative to other participants. That is:*

$$\begin{aligned} \partial^2\mu_j/\partial c_j\partial f_j &> 0 \; \forall\theta_i \text{ if } \omega_i > 0 \\ \partial^2\mu_j/\partial c_j\partial f_j &< 0 \; \forall\theta_i \text{ if } \omega_i < 0 \end{aligned} \tag{4}$$

In particular, if they believe women and men have the same ability, then gender of the person who corrects them does not matter. That is:

$$\partial^2\mu_j/\partial c_j\partial f_j = 0 \; \forall\theta_i \text{ if } \omega_i = 0 \tag{5}$$

I consider deviations from these propositions are evidence of non-rationality.

5 Response to corrections

In this section, I document evidence that participants – both women and men – understand the notion of good and bad moves. Thus, they are more willing to collaborate with someone who contributed more to the puzzle. This is in line with what one would expect and validates my experimental design.

However, after controlling for that person's contribution, participants are less willing to collaborate with a person who corrected their move, even if that person makes good corrections, which is inefficient and seems to indicate deviation from the rational agent's benchmark in Proposition 1. However, it is unclear whether it is evidence of deviation from rationality; I will return to this point later in section 6.

Estimating equation I estimate the following model with OLS.

$$Select_{ij} = \beta_1 CorrectedGood_{ij} + \beta_2 CorrectedBad_{ij} + \beta_3 Female_j + \delta Contribution_j + \mu_i + \epsilon_{ij} \tag{6}$$

where each variable is defined as follows:

- $Select_{ij} \in \{0, 1\}$: an indicator variable equals 1 if i selects j as their collaborator, 0 otherwise.
- $CorrectedGood_{ij} \in \{0, 1\}$: an indicator variable equals 1 if j corrected i and moved the puzzle closer to the solution, 0 otherwise.
- $CorrectedBad_{ij} \in \{0, 1\}$: an indicator variable equals 1 if j corrected i and moved the puzzle far away from the solution, 0 otherwise.
- $Female_j \in \{0, 1\}$: an indicator variable equals 1 if j is female, 0 otherwise.
- $Contribution_j \in \mathbb{Z}$: js contribution to a puzzle played with i.
- ϵ_{ij}: omitted factors that affect i's likelihood to select j as their collaborator.

and $\mu_i \equiv \sum_{k=1}^{N} \mu^k \mathbb{1}[i = k]$ is individual fixed effects, where N is the total number of participants in the sample and $\mathbb{1}$ is the indicator variable. Standard errors are clustered at the individual level.[24]

The key identification assumption is that $Contribution_j$ fully captures j's ability perceived by i through j's puzzle moves (not true ability). This assumption is reasonable if we think participants' willingness to collaborate is increasing in the partner's contribution to the puzzle, which is consistent with that participants can partially observe their partners' ability and their expected utility is increasing in their payoff.

The identification discussion deserves a further elaboration: by random pairing of participants, the paired participant's gender is exogenous to participant's unobservables. However, correction is not exogenous for two reasons: (i) correction can be correlated with the paired participant's ability and paired participant's ability can affect participant's collaborator selection; (ii) There is an effect similar to the reflection effect: participant's puzzle behavior affects the paired participant's behavior and vice versa. For example, participant's meanness can increase the paired participant's correction and can also affect their collaborator selection. The identification assumption concerns the former point. To address the latter point, I add individual fixed effects.

Results Table 3 presents the regression results of equation 6. Columns 1-4 include all participants' willingness to collaborate but column 1 excludes partner's contribution and individual fixed effects

24. This is because the treatment unit is i. Although the same participant appears twice (once as i and once as j), j is passive in collaborator selection.

Table 3: Response to corrections

Dependent variable:	Willing to collaborate (yes=1, no=0)									
Sample:	All					Female			Male	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Good correction	-0.208***	-0.238***		-0.204***	-0.269***		-0.229***	-0.197***		-0.168***
	(0.028)	(0.030)		(0.024)	(0.043)		(0.033)	(0.040)		(0.036)
Bad correction	-0.518***	-0.508***		-0.100***	-0.550***		-0.172***	-0.457***		-0.011
	(0.031)	(0.034)		(0.036)	(0.044)		(0.047)	(0.050)		(0.052)
Any correction			-0.198***			-0.237***			-0.152***	
			(0.022)			(0.030)			(0.031)	
Female partner	-0.003	-0.001	0.008	0.009	-0.009	0.002	0.004	0.007	0.016	0.016
	(0.016)	(0.017)	(0.014)	(0.014)	(0.021)	(0.018)	(0.018)	(0.026)	(0.021)	(0.021)
Partner's contribution			0.083***	0.084***		0.090***	0.089***		0.077***	0.080***
			(0.003)	(0.003)		(0.004)	(0.004)		(0.003)	(0.004)
Individual FE		✓	✓	✓	✓	✓	✓	✓	✓	✓
P-value: Good correction =Bad correction	0.000	0.000		0.020	0.000		0.347	0.000		0.016
Baseline mean	0.780	0.780	0.780	0.780	0.780	0.780	0.780	0.778	0.778	0.778
Baseline SD	0.414	0.414	0.414	0.414	0.414	0.414	0.414	0.416	0.416	0.416
Adj. R-squared	0.104	0.100	0.334	0.335	0.111	0.365	0.369	0.090	0.306	0.306
Observations	3180	3180	3180	3180	1670	1670	1670	1510	1510	1510
Individuals	464	464	464	464	244	244	244	220	220	220

Notes: This table presents the regression results of equation 6. Columns 1-4 include all participants' willingness to collaborate but column 1 excludes partner's contribution and individual fixed effects and column 2 partner's contribution. Column 3 combines good and bad correction as a single dummy variable. Columns 5-7 present the corresponding results for women and columns 8-10 for men. Baseline mean and standard deviation are willingness to collaborate with partners who do not make any corrections. Standard errors in parentheses are clustered at the individual level. Significance levels: * 10%, ** 5%, and *** 1%.

and column 2 partner's contribution. Column 3 combines good and bad correction as a single dummy variable. Columns 5-7 present the corresponding results for women and columns 8-10 for men.

Column 1 shows that when we do not control for between-participants variation, the coefficient estimate on good correction is underestimated. Column 2 shows that when we do not control for partner's contribution, the coefficient estimate on bad correction is negative and very large: the point estimate is -0.508 (p-value < 0.01); that is, participants are 50.8% less willing to collaborate with partners who made a bad correction, a correction that moved the puzzle far away from the solution. Indeed, these coefficient estimates are more negative than the coefficient estimates on good corrections: 0.271 more negative (p-value < 0.01). This is true when we separately examine women (column 5, 0.281 with p-value < 0.01) and men (column 8, 0.281 with p-value < 0.01).

Corroborating this, looking at column 3, the coefficient estimate on the partner's contribution is positive and quantitatively and statistically highly significant and is 0.083 (p-value < 0.01). This suggests that participants are 8.3% more willing to collaborate with partners who make one more good move. This is true for women (column 6, 0.090 with p-value < 0.01) and men (column 9, 0.077 with p-value < 0.01). This is evidence that my experimental design is valid: participants correctly understand the notion of good and bad moves and that participants are more willing to collaborate with partners who contributed more.

The coefficient estimate on any correction in column 3 is negative and quantitatively and statistically highly significant and is -0.198 (p-value < 0.01). This suggests that people are 19.8%

less willing to collaborate with those who made a correction(s). To offset this effect, a partner's contribution has to increase by 0.79 standard deviation.[25] The corresponding coefficient estimate for women is -0.237 (column 6, p-value < 0.01) and -0.152 for men (column 9, p-value < 0.01). Thus, participants are less willing to collaborate with a person who corrected their move.

This is not a problem if participants are more willing to collaborate with a person who made a good correction and less willing to collaborate with a person who made a bad correction. However, this is not the case: the coefficient estimate on good correction in column 4 is still negative and is -0.204 (p-value < 0.01). This suggests that people are less willing to collaborate even with those who made a good correction(s). The corresponding coefficient estimate for women is -0.229 (column 7, p-value < 0.01) and -0.168 for men (column 10, p-value < 0.01).

The coefficient estimate on bad correction in columns 4 is also negative and quantitatively and statistically significant and is -0.100 (p-value < 0.01). However, the magnitude is smaller than the coefficient estimate on good correction: the difference is -0.104 (p-value < 0.05). This is mainly driven by men: the corresponding coefficient estimate for women is -0.172 (column 6, p-value $<$ 0.01) but is -0.011 (p-value > 0.10) for men.

These behaviors are inefficient. They also seem to indicate deviation from the rational agent's benchmark in Proposition 1. However, we cannot say anything definitive about rationality because response to corrections depends on the belief about people's own ability relative to partners and people are in general overconfident albeit that men are more overconfident (Croson and Gneezy 2009). Thus, these behaviors are not irrational if participants believe they are better at the puzzle than others. I will come back to this point in section 6.

6 Who respond negatively to corrections?

If only low ability people respond negatively to corrections, then correcting colleagues may not be very costly, because collaborations with those people will likely result in low quality outcomes. Also, it was inconclusive whether people's negative reaction to corrections documented in section 5 is irrational. However, if it is rational, people with higher puzzle solving ability – as measured in the part 1 individual practice stage – should respond less negatively to good corrections and more negatively to bad corrections because they are better able to distinguish good and bad corrections than people with lower puzzle solving ability as in Proposition 2.

In this section, I document that even high ability people respond negatively to good corrections with men responding more negatively, which indicates that (i) correcting colleagues is really costly both for individuals and groups and that (ii) negative response to corrections is irrational.

25. The number is calculated as follows: $\hat{\beta}_{Partner's\ contribution} \times SD_{Partner's\ contribution} \times x = |\hat{\beta}_{Any\ correction}| \Rightarrow x = |\hat{\beta}_{Any\ correction}|/(\hat{\beta}_{Partner's\ contribution} \times SD_{Partner's\ contribution}) = 0.198/(0.09 \times 2.8) \approx 0.79$. $SD_{Partner's\ contribution} = 2.8$ is from panel B of Table 2 and is an arithmetic average of 2.73 for partners faced by women 2.87 for and partners faced by men: (2.73+2.87)/2=2.80.

Estimating equation I estimate the following model with OLS.

$$\begin{aligned} Select_{ij} =& \beta_1 CorrectedGood_{ij} + \beta_2 CorrectedBad_{ij} + \beta_3 Female_j \\ &+ \beta_4 CorrectedGood_{ij} \times HighAbility_i + \beta_5 CorrectedBad_{ij} \times HighAbility_i \\ &+ \delta_1 Contribution_j + \delta_2 Contribution_j \times HighAbility_i + \mu_i + \epsilon_{ij} \end{aligned} \tag{7}$$

where each variable is defined as follows:

- $HighAbility_i \in \{0, 1\}$: an indicator variable equals 1 if i solved the above-median number of puzzles in part 1 in a session they have participated, 0 otherwise.

Other variables are as defined in equations 6.

Panel B of Figure 4 shows distribution of contribution of high ability people is indeed less dispersed, consistent with that I matched up high ability people with high ability people – and low ability people with low ability people – the collaborator selection stage in part 2.

Results Table 4 presents the regression results of equation 8. As Table 3, columns 1-2 include all participants' willingness to collaborate. columns 3-4 the corresponding results for women and columns 5-6 for men.

In column 1, the coefficient estimate on the interaction between any correction and high ability is negative and statistically significant (p-value < 0.05). This effect mainly comes from men: the corresponding coefficient estimate for women (column 3) is less negative and statistically insignificant, but is more for men (column 5, p-value < 0.05). Thus, high ability people, in particular men, dislike receiving corrections more than low ability people.

It is not a problem if this result is coming from high ability people responding less negatively or even positively to good corrections and more negatively to bad corrections. However, this is not the case: in column 2, the coefficient estimate on the interaction between good correction and high ability is negative (p-value < 0.05). This effect comes from both women and men, with the effect on men being stronger: the corresponding coefficient estimate for women (column 4) is negative albeit less so and statistically insignificant, and is more negative and statistically significant (p-value $<$ 0.05) for women (in column 6).

The coefficient estimate on the interaction between bad correction and high ability in column 2 is almost zero. The corresponding coefficient estimate is positive for women (column 4) and negative for men (column 6), although they are both statistically insignificant.

Thus, even high ability people respond negatively to good corrections with men responding more negatively. This suggests that correcting colleagues is really costly both for individuals and groups because even high ability people are less willing to collaborate with someone who corrects their mistakes.

This result also suggests that negative reaction to corrections is likely to be irrational: as discussed at the beginning of this section, high ability people should be able to distinguish good and bad corrections and should respond less negatively to good corrections and more negatively to bad corrections than low ability people as the rational agent benchmark in Proposition 2 suggests.

Table 4: Response to corrections of high vs. low ability people

Dependent variable:	Willing to collaborate (yes=1, no=0)					
Sample:	All		Female		Male	
	(1)	(2)	(3)	(4)	(5)	(6)
Good correction		-0.155***		-0.208***		-0.107***
		(0.030)		(0.042)		(0.041)
Bad correction		-0.100**		-0.201***		0.005
		(0.047)		(0.064)		(0.063)
Any correction	-0.153***		-0.213***		-0.096**	
	(0.028)		(0.041)		(0.037)	
Female partner	0.008	0.009	0.002	0.002	0.015	0.014
	(0.014)	(0.014)	(0.018)	(0.018)	(0.021)	(0.021)
Partner's contribution	0.084***	0.084***	0.090***	0.089***	0.079***	0.082***
	(0.003)	(0.004)	(0.005)	(0.005)	(0.004)	(0.004)
Good correction x High ability		-0.118**		-0.048		-0.180**
		(0.050)		(0.066)		(0.075)
Bad correction x High ability		0.000		0.074		-0.061
		(0.072)		(0.095)		(0.109)
Any correction x High ability	-0.108**		-0.051		-0.152**	
	(0.044)		(0.061)		(0.064)	
Partner's contribution x High ability	-0.002	-0.001	-0.002	-0.001	-0.004	-0.003
	(0.005)	(0.005)	(0.007)	(0.007)	(0.007)	(0.008)
Individual FE	✓	✓	✓	✓	✓	✓
Baseline mean	0.780	0.780	0.783	0.783	0.778	0.778
Baseline SD	0.414	0.414	0.413	0.413	0.416	0.416
Adj. R-squared	0.335	0.336	0.365	0.368	0.308	0.308
Observations	3180	3180	1670	1670	1510	1510
Individuals	464	464	244	244	220	220

Notes: This table presents the regression results of equation 8. Columns 1-2 include all participants' willingness to collaborate. Columns 3-4 present the corresponding results for women and columns 5-6 for men. Baseline mean and standard deviation are willingness to collaborate with partners who do not make any corrections. Standard errors in parentheses are clustered at the individual level. Significance levels: * 10%, ** 5%, and *** 1%.

However, what we see here is the opposite. At least, it is inconsistent.

7 Do women's corrections receive stronger negative reactions?

In this section, I document that people – either men or women – do not underestimate women's contribution, which suggests that their prior about women's ability to solve the puzzle is neither higher nor lower than men. However, I also document suggestive evidence that women's good corrections may receive stronger negative reactions by men, which is inconsistent with Proposition 3. This is a problem for women's career success because men are still majority in top positions in society and hence missing collaboration opportunities with them could be detrimental.

Estimating equation I estimate the following model with OLS.

$$\begin{aligned} Select_{ij} = & \beta_1 CorrectedGood_{ij} + \beta_2 CorrectedBad_{ij} + \beta_3 Female_j \\ & + \beta_4 CorrectedGood_{ij} \times Female_j + \beta_5 CorrectedBad_{ij} \times Female_j \\ & + \delta_1 Contribution_j + \delta_2 Contribution_j \times Female_j + \mu_i + \epsilon_{ij} \end{aligned} \tag{8}$$

Where each variable is defined as in equation 6.

Table 5: Response to corrections made by women vs. men

Dependent variable:	Willing to collaborate (yes=1, no=0)					
Sample:	All		Female		Male	
	(1)	(2)	(3)	(4)	(5)	(6)
Good correction		-0.187***		-0.248***		-0.104*
		(0.035)		(0.045)		(0.053)
Bad correction		-0.176***		-0.218***		-0.104
		(0.051)		(0.064)		(0.076)
Any correction	-0.203***		-0.260***		-0.125***	
	(0.031)		(0.042)		(0.045)	
Female partner	0.013	0.001	-0.001	-0.002	0.026	0.003
	(0.022)	(0.022)	(0.032)	(0.032)	(0.029)	(0.030)
Partner's contribution	0.084***	0.083***	0.090***	0.089***	0.078***	0.077***
	(0.004)	(0.004)	(0.006)	(0.006)	(0.005)	(0.006)
Good correction x Female partner		-0.035		0.035		-0.119*
		(0.044)		(0.057)		(0.067)
Bad correction x Female partner		0.144**		0.090		0.168
		(0.070)		(0.093)		(0.102)
Any correction x Female partner	0.009		0.047		-0.051	
	(0.041)		(0.056)		(0.059)	
Partner's contribution x Female partner	-0.002	0.002	-0.001	-0.001	-0.001	0.006
	(0.005)	(0.005)	(0.008)	(0.008)	(0.007)	(0.007)
Individual FE	✓	✓	✓	✓	✓	✓
Baseline mean	0.780	0.780	0.783	0.783	0.778	0.778
Baseline SD	0.414	0.414	0.413	0.413	0.416	0.416
Adj. R-squared	0.333	0.336	0.365	0.369	0.305	0.307
Observations	3180	3180	1670	1670	1510	1510
Individuals	464	464	244	244	220	220

Notes: This table presents the regression results of equation 8. Columns 1-2 include all participants' willingness to collaborate. Columns 3-4 present the corresponding results for women and columns 5-6 for men. Baseline mean and standard deviation are willingness to collaborate with partners who do not make any corrections. Standard errors in parentheses are clustered at the individual level. Significance levels: * 10%, ** 5%, and *** 1%.

Results Table 5 presents the regression results of equation 8. As Table 3, columns 1-2 include all participants' willingness to collaborate, columns 3-4 present the corresponding results for women and columns 5-6 for men.

Looking at column 1, the coefficient estimate on the interaction between partner's contribution and female partner is almost 0 and statistically insignificant. Column 3 shows this is true for women and column 5 for men. These suggest that people – both women and men – do not underestimate

women's contribution when selecting a collaborator. In other words, people correctly believe that women and men are equally good at solving the puzzle.

In column 1, the coefficient estimates on the interaction between any correction and female partner is close to 0 and statistically insignificant. However, women and men respond differently: the corresponding coefficient estimate is positive for women (column 3) but negative for men (column 5), although they are statistically insignificant.

Column 2 splits any correction into good and bad correction and shows asymmetric response: the coefficient estimates on the interaction between good correction and female partner is negative although statistically insignificant, but the coefficient estimates on the interaction between female partner and bad correction is positive (p-value < 0.05).

The negative coefficient estimate on the interaction between good correction and female partner mainly comes from men: looking at column 6, the corresponding coefficient estimate for men is -0.119 and marginally significant (p-value < 0.10), while for women it is 0.035 although statistically insignificant (column 4). On the other hand, the positive coefficient estimate on the interaction between female partner and bad correction comes from both women and men: the corresponding coefficient estimate is 0.090 for women (column 4) and 0.168 for men (column 6), although neither of them is statistically significant. Together with the evidence that men believe women are equally good at solving the puzzle as men, this is inconsistent with Proposition 3.

Men's less negative – or even positive – response to women's bad correction is a bit puzzling. One explanation is that men do not like to be corrected for their mistakes by women, but they are okay that women make mistakes.

Although only statistically marginally significant, men's negative reaction to women's good correction is worrying as men are still the majority in top positions both in industry (Bertrand 2018) and academia (Lundberg and Stearns 2019); missing collaboration opportunities with them could be detrimental for women's career.

8 Robustness checks

Excluding unsolved puzzles Whether participants can solve a puzzle is an outcome of a particular pairing that is random. However, "a good move is only preferable if you are playing with a partner who is also trying to solve the puzzle" (Isaksson 2018, p. 25). If a participant is not trying to solve the puzzle, then a pair is unlikely to solve the puzzle and good and bad corrections may not be meaningful.

Excluding rounds 6 and 7 Remember that each participant plays the puzzle for seven rounds and variables unaffected by treatment (interactions within a randomly-formed pair) must be balanced. Figure 5 plots average partner gender balance (fraction of female partners, panel A) and puzzle outcomes (panels B-H) across seven rounds along with their 95% confidence intervals, separately for female (blue) and male participants (green). We see that in rounds 6 and 7, participants are less

willing to collaborate, experience more corrections, and less likely to solve the puzzle. Although they are all outcomes of a particular pair so can just be correlations, one may wonder whether rounds 6 and 7 are driving the results.

Excluding puzzles where good and bad corrections occurred There are 495 puzzles in which at least one correction occurred, of which 325 puzzles experienced good corrections only, 110 puzzles bad corrections only, and 60 puzzles experienced both good and bad corrections. In these 60 puzzles, it is unclear which corrections – good or bad – dominated people's mind in determining whether to collaborate with a paired person.

To address these concerns, I re-estimate equations 6, 7, and 8, and plot the coefficient estimates and 95% confidence intervals of the main coefficients of interest in Figures 6, 7, and 8, respectively, with solved puzzles only (green dots and lines), with rounds 1-5 only (red dots and lines), and with puzzles where only good or bad corrections occurred (purple dots and lines). As a reference, I also plot the coefficient estimates and 95% confidence intervals with the main sample used in Tables 3, 4, and 5 (blue dots and lines). All estimates are from the full models (columns 4, 7, and 10 for Table 3 and columns 2, 4, and 6 for Tables 4 and 5).

The main coefficients of interest for equation 6 are good and bad corrections. Looking at Figure 6, we see that most coefficient estimates are close to the main estimates. The estimates are more negative for good correction when the sample is limited to solved puzzles only, but they are more in line with the main findings.

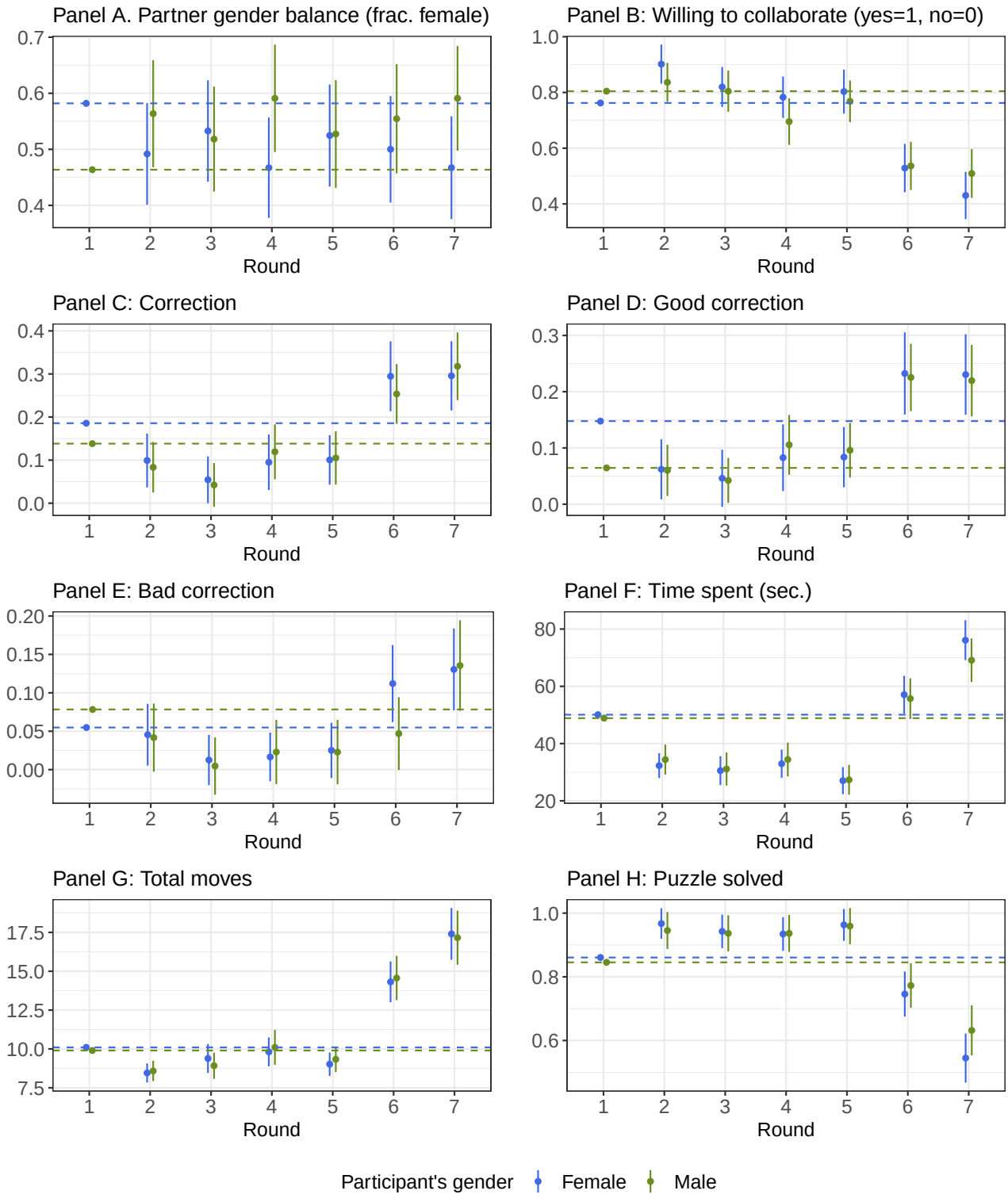
The main coefficients of interest for equation 7 are the interaction between good correction and high ability and between bad correction and high ability. Looking at Figure 7, we again see most the coefficient estimates are close to the main estimates.

The main coefficients of interest for equation 8 are the interaction between good correction and female partner and between bad correction and female partner. Looking at Figure 8, we again see most the coefficient estimates are close to the main estimates. Again, the estimates with solved puzzles only present somewhat different evidence; in particular, response to good corrections by female partners is negative although statistically insignificant for women and positive for men. However, both estimates are very close to 0 and do not contradict with that the evidence that men react more negatively to women's good correction is only suggestive.

9 Conclusion

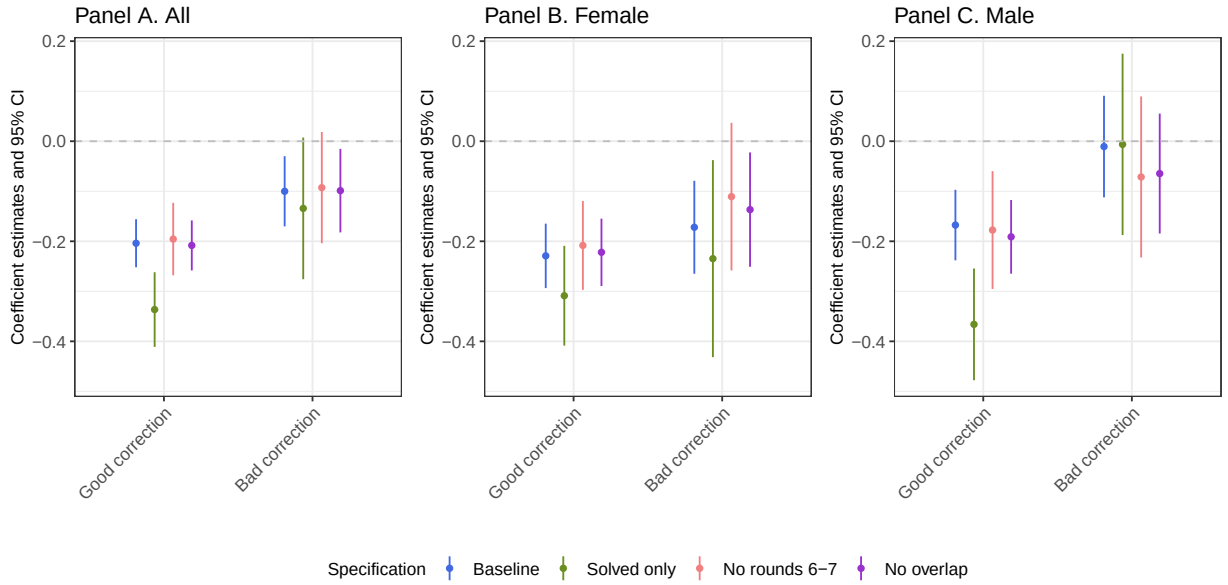
This paper demonstrates that people, including those with high productivity, are less willing to collaborate with a person who has corrected them even if the correction improves group performance. I also find suggestive evidence that men respond more negatively to women's corrections that improves group performance, presumably because men do not like to be corrected for their mistakes by women. Thus, dislike to be corrected distorts the optimal selection of talents and penalizes those who correct others' mistakes, and the distortion may be stronger when women correct men.

Figure 5: Balance across rounds



Notes: This figure shows point estimates and 95% confidence intervals of βs from the following OLS regression with gender balance (female dummy) and different puzzle outcomes separately for female (blue) and male participants (green): $y_{ij} = \beta_1 + \sum_{k=2}^{7} \beta_k \mathbb{1}[t_{ij} = k] + \epsilon_{ij}$, where $t_{ij} \in \{1, 2, 3, 4, 5, 6, 7\}$ is the puzzle round in which i and j are playing, $\mathbb{1}$ is an indicator variable, and y_{ij} is dependent variable indicated in each panel. I add the estimate of β_1 to estimates of β_2-β_7 to make the figure easier to look at. Standard errors are clustered at the individual level.

Figure 6: Response to corrections: Robustness



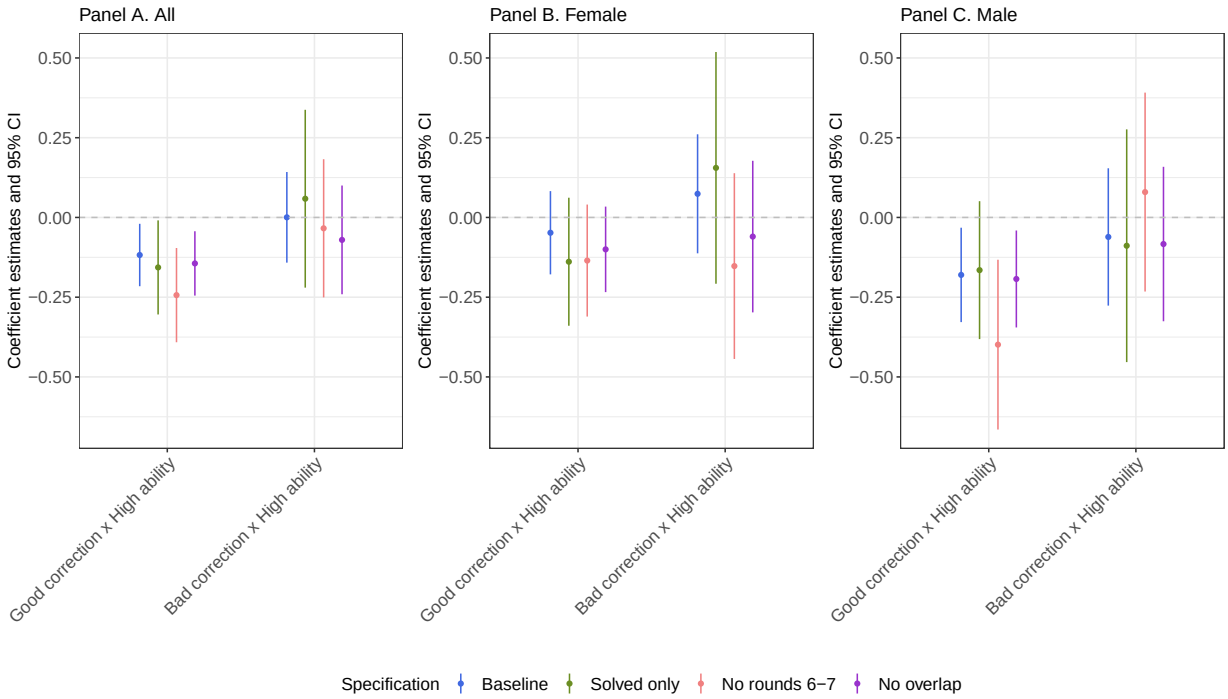
Notes: This figure plots the coefficient estimates and 95% confidence intervals of columns 4, 7, and 10 of Table 3 with solved puzzles only (the green dots and lines), with rounds 1-5 only (the red dots and lines), and with puzzles where only good or bad corrections occurred (the purple dots and lines). The blue dots and lines are the corresponding baseline estimates. They show that the findings in Table 3 are robust to limiting samples in these ways.

While a laboratory setting is different from the real-world, my findings are likely to be a lower bound because of the following three reasons. First, there is no reputation cost in my experiment: being corrected is not observed by others, unlike in the real-world. Second, the emotional stake is much smaller in my experiment: the puzzle ability is not informative of the ability relevant for their work or study – it is not something people have been devoting much of their time to, such as university exams, academic research, or corporate investment projects. Third, participants are equal in my experiment; in the real-world, there are sometimes senior-junior relationships, and corrections by junior people may induce stronger negative reactions. Thus, introducing reputation cost, using tasks that are more related to one's real-world ability, and having age variation would be interesting extensions of this paper.

But there are two limitations. The first is that participants are strangers to each other in my experiments while people know each other in the real-world. Thus, it is possible that repeated interactions would mitigate people's negative response to corrections (but they may also magnify the negative response due to rivalry, failure to build a good rapport, etc.). The second limitation is that most participants are bachelor or master students who are supposed to have a weaker gender bias. The first point relates to the takeaway of my results: it would be worth investigating whether good workplace climate mitigates negative reaction to corrections. The second point relates to external validity: women's corrections may receive stronger and more robust negative reactions in the real-world where people are older.

Finally, my experiment is not designed to investigate the underlying mechanism, but the results

Figure 7: Response to corrections made of high vs. low ability people: Robustness

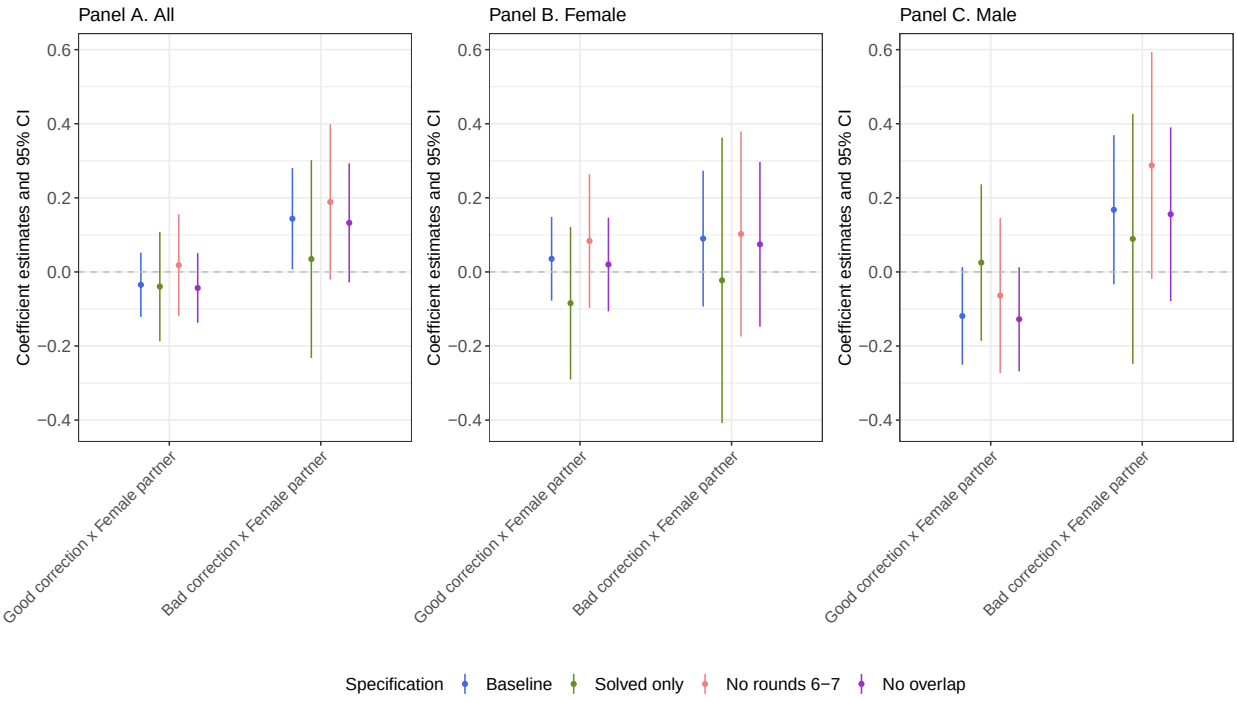


Notes: This figure plots the coefficient estimates and 95% confidence intervals of columns 2, 4, and 6 of Table 4 with solved puzzles only (the green dots and lines), with rounds 1-5 only (the red dots and lines), and with puzzles where only good or bad corrections occurred (the purple dots and lines). The blue dots and lines are the corresponding baseline estimates. They show that the findings in Table 4 are robust to limiting samples in these ways.

are consistent with the findings of the literature on self-image concerns and information avoidance. For example, Kszegi (2006) finds that people avoid a difficult task when it reveals their ability. Corroborating this, Castagnetti and Schmacker (2021) find people select information less informative about their ability and Ewers and Zimmermann (2015) find people exaggerate their ability when others observe it even at the cost of reducing their payoff. There is also evidence that people respond differently to women's and men's feedback (Sinclair and Kunda 2000). In light of this literature, my results can be interpreted as follows: receiving good corrections is negative feedback and accepting them as correct damages people's self-image.[26] Investigating how people's information avoidance interacts with workplace climate – and whether gender matters – would be an interesting future research.

26. Which means θ_i in the theoretical model in section 4 (equation 1) is not exogenous.

Figure 8: Response to corrections made by women vs. men: Robustness



Notes: This figure plots the coefficient estimates and 95% confidence intervals of columns 2, 4, and 6 of Table 5 with solved puzzles only (the green dots and lines), with rounds 1-5 only (the red dots and lines), and with puzzles where only good or bad corrections occurred (the purple dots and lines). The blue dots and lines are the corresponding baseline estimates. They show that the findings in Table 5 are robust to limiting samples in these ways.

References

Abelson, Robert P. 1986. "Beliefs Are Like Possessions." *Journal for the Theory of Social Behaviour* 16 (3): 223–250.

Alan, Sule, Gozde Corekcioglu, and Matthias Sutter. 2021. *Improving Workplace Climate in Large Corporations: A Clustered Randomized Intervention.* Working Paper.

Arechar, Antonio A., Simon Gächter, and Lucas Molleman. 2018. "Conducting Interactive Experiments Online." *Experimental Economics* 21 (1): 99–131.

Bertrand, Marianne. 2018. "Coase Lecture – The Glass Ceiling." *Economica* 85 (338): 205–231.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. 2019. "Beliefs about Gender." *American Economic Review* 109 (3): 739–773.

Carrell, Scott E., Marianne E. Page, and James E. West. 2010. "Sex and Science: How Professor Gender Perpetuates the Gender Gap." *The Quarterly Journal of Economics* 125 (3): 1101–1144.

Castagnetti, Alessandro, and Renke Schmacker. 2021. *Protecting the Ego: Motivated Information Selection and Updating.* Working Paper.

Chen, Daniel L., Martin Schonger, and Chris Wickens. 2016. "oTree–An Open-Source Platform for Laboratory, Online, and Field Experiments." *Journal of Behavioral and Experimental Finance* 9:88–97.

Coffman, Katherine, Clio Bryant Flikkema, and Olga Shurchkov. 2021. "Gender Stereotypes in Deliberation and Team Decisions." *Games and Economic Behavior* 129:329–349.

Coffman, Katherine Baldiga. 2014. "Evidence on Self-Stereotyping and the Contribution of Ideas." *The Quarterly Journal of Economics* 129 (4): 1625–1660.

Croson, Rachel, and Uri Gneezy. 2009. "Gender Differences in Preferences." *Journal of Economic Literature* 47 (2): 448–474.

Dupas, Pascaline, Alicia Sasser Modestino, Muriel Niederle, Justin Wolfers, and Seminar Dynamics Collective. 2021. *Gender and the Dynamics of Economics Seminars.* Working Paper.

Edmans, Alex. 2011. "Does the Stock Market Fully Value Intangibles? Employee Satisfaction and Equity Prices." *Journal of Financial Economics* 101 (3): 621–640.

Egan, Mark, Gregor Matvos, and Amit Seru. 2021. "When Harry Fired Sally: The Double Standard in Punishing Misconduct." *Journal of Political Economy.*

Ewers, Mara, and Florian Zimmermann. 2015. "Image and Misreporting." *Journal of the European Economic Association* 13 (2): 363–380.

Fisman, Raymond, Sheena S. Iyengar, Emir Kamenica, and Itamar Simonson. 2006. "Gender Differences in Mate Selection: Evidence from a Speed Dating Experiment." *The Quarterly Journal of Economics* 121 (2): 673–697.

———. 2008. "Racial Preferences in Dating." *The Review of Economic Studies* 75 (1): 117–132.

Gallus, Jana, and Emma Heikensten. 2019. *Shine a Light on the Bright: The Effect of Awards on Confidence to Speak up in Gender-Typed Knowledge Work.* Working Paper.

Glick, Peter, and Susan T. Fiske. 1996. "The Ambivalent Sexism Inventory: Differentiating Hostile and Benevolent Sexism." *Journal of Personality and Social Psychology* (US) 70 (3): 491–512.

Goeschl, Timo, Marcel Oestreich, and Alice Soldà. 2021. *Competitive vs. Random Audit Mechanisms in Environmental Regulation: Emissions, Self-Reporting, and the Role of Peer Information.* Working Paper 0699. University of Heidelberg, Department of Economics.

Golman, Russell, David Hagmann, and George Loewenstein. 2017. "Information Avoidance." *Journal of Economic Literature* 55 (1): 96–135.

Greiner, Ben. 2015. "Subject Pool Recruitment Procedures: Organizing Experiments with ORSEE." *Journal of the Economic Science Association* 1 (1): 114–125.

Guiso, Luigi, Paola Sapienza, and Luigi Zingales. 2015. "The Value of Corporate Culture." *Journal of Financial Economics,* NBER Conference on the Causes and Consequences of Corporate Culture, 117 (1): 60–76.

Guo, Joyce, and María P. Recalde. 2020. *Overriding in Teams: The Role of Beliefs, Social Image, and Gender.* Working Paper.

Harrison, Glenn W., and John A. List. 2004. "Field Experiments." *Journal of Economic Literature* 42 (4): 1009–1055.

Isaksson, Siri. 2018. *It Takes Two: Gender Differences in Group Work.* Working Paper.

Jones, Benjamin F. 2021. "The Rise of Research Teams: Benefits and Costs in Economics." *Journal of Economic Perspectives* 35 (2): 191–216.

Kszegi, Botond. 2006. "Ego Utility, Overconfidence, and Task Choice." *Journal of the European Economic Association* 4 (4): 673–707.

Lazear, Edward P., and Kathryn L. Shaw. 2007. "Personnel Economics: The Economist's View of Human Resources." *Journal of Economic Perspectives* 21 (4): 91–114.

Lundberg, Shelly, and Jenna Stearns. 2019. "Women in Economics: Stalled Progress." *Journal of Economic Perspectives* 33 (1): 3–22.

Sarsons, Heather. 2019. *Interpreting Signals in the Labor Market: Evidence from Medical Referrals.* Working Paper.

Sinclair, Lisa, and Ziva Kunda. 2000. "Motivated Stereotyping of Women: Shes Fine If She Praised Me but Incompetent If She Criticized Me." *Personality and Social Psychology Bulletin* 26 (11): 1329–1342.

Stoddard, Olga, Christopher F. Karpowitz, and Jessica Preece. 2020. *Strength in Numbers: A Field Experiment in Gender, Influence, and Group Dynamics.* Working Paper.

Wuchty, Stefan, Benjamin F. Jones, and Brian Uzzi. 2007. "The Increasing Dominance of Teams in Production of Knowledge." *Science* 316 (5827): 1036–1039. pmid: 17431139.

Appendix

Appendix A Advantages and drawbacks of quasi-laboratory format

The quasi-laboratory experimental format I use in this paper has advantages and drawbacks over the standard online and physical laboratory format.

Over standard online format The main advantage over standard online experimental format is that we can mostly avoid attrition, which is the main problem of online interactive experiments (Arechar, Gächter, and Molleman 2018). The reason is that compared to platforms such as MTurk and Prolific where participants' identity is fully anonymous by their rule, we have participants' personal information and participants know it as we recruit them from our standard laboratory subject pool. Also, they are connected to us via Zoom throughout the experiment. In my experiment, I experienced no participant attrition.

Another advantage is that we can fully control over who will participate in the experiment. For instance, we can screen out participants who have participated in particular kinds of experiments, such as experiments involving deception which is another problem of online experiments (Arechar, Gächter, and Molleman 2018). In my case, I have excluded participated who have participated in gender-related experiments in the past. This allows us to collect cleaner data.

The key drawback is the difficulty of collecting a large number of observations. Unlike MTurk or Prolific, the experimenter has to be present and respond to participants, if necessary, throughout the experiment. We could recruit a large number of participants at once, for example several hundreds, but it weakens the connection between the experimenter and the participants and can induce attrition.

Over physical laboratory format The main advantage over physical laboratory format is logistical convenience both for the experimenter and the participants: we can run and join experiments from our offices or home as long as we have a computer and an internet connection. It primarily benefited me to comply with the COVID-19 precautions. However, it also means that we can run laboratory experiments even if we do not have a physical laboratory in our university, for example in universities in low-income countries, as long as we set up ORSEE (Greiner 2015) or other subject management system, many of which are free.

Another advantage is that since participants can join the experiments from anywhere in the world, we can potentially run experiments with non-standard subjects or what Harrison and List (2004) call artefactual field experiments. For instance, non-student subjects or subjects in other countries. Although there can be regulation issues we have to overcome, it increases the kind of questions we can answer.

There are already a few studies that use a quasi-laboratory format, for example, Goeschl, Oestreich, and Soldà (2021).

Appendix B Heterogeneity by gender bias

I have pre-specified in the pre-analysis plan that investigated heterogeneous response to women's corrections by the degree of gender bias measured by the six hostile and benevolent sexism questions used in Stoddard, Karpowitz, and Preece (2020), selected from Glick and Fiske (1996). Although I could not detect a meaningful variation, I present the results here.

I estimate the following model with OLS.

$$\begin{aligned} Select_{ij} = & \beta_1 CorrectedGood_{ij} + \beta_2 CorrectedBad_{ij} + \beta_3 Female_j \\ & + \beta_4 CorrectedGood_{ij} \times HighBias_i + \beta_5 CorrectedBad_{ij} \times HighBias_i \\ & + \delta_1 Contribution_j + \delta_2 Contribution_j \times HighBias_i + \mu_i + \epsilon_{ij} \end{aligned} \tag{B1}$$

where each variable is defined as follows:

- $HighBias_i \in \{0, 1\}$: an indicator variable equals 1 if i's gender bias score from the six hostile and benevolent sexism questions is above median among participants with the same gender (female or male), 0 otherwise.

Other variables are as defined in equations 6.

Table B1 presents the regression results of equation B1. As Table 3, columns 1-2 include all participants' willingness to collaborate. columns 3-4 the corresponding results for women and columns 5-6 for men.

In columns 1, 3, and 5, the coefficient estimate on the interaction among any correction, female partner, and high bias is negative but statistically insignificant. Also, in column 2, 4, and 6, the coefficient estimate on the interaction between good correction, female partner, and high bias as well as on the interaction between bad correction, female partner, and high bias are mostly negative but statistically insignificant. Thus, while the gender bias measure may detect some of the gender bias of participants, it is not variable enough to capture any meaningful heterogeneity among participants.

Appendix C Results with the original contribution measure

In the main text, I changed the definition of contribution from the one in the pre-analysis plan because there was truncation in the original contribution measure in more than 10% of the puzzle. Nonetheless, the same results hold when I use the original contribution measure, reported in Tables C1, C2, C3, and C4. Although the original measure is relative to one's pair while the measure I use in this paper is absolute, whether a measure is relative or absolute does not matter because I add individual fixed effects.

Table B1: Response to corrections made by women vs. men: Heterogeneity by gender bias

Dependent variable:	Willing to collaborate (yes=1, no=0)					
Sample:	All		Female		Male	
	(1)	(2)	(3)	(4)	(5)	(6)
Good correction		-0.180***		-0.260***		-0.066
		(0.046)		(0.058)		(0.071)
Bad correction		-0.196***		-0.185**		-0.201*
		(0.071)		(0.094)		(0.104)
Any correction	-0.215***		-0.267***		-0.137**	
	(0.041)		(0.053)		(0.065)	
Female partner	0.023	0.004	0.022	0.014	0.025	-0.006
	(0.029)	(0.029)	(0.039)	(0.040)	(0.041)	(0.042)
Partner's contribution	0.088***	0.087***	0.094***	0.094***	0.082***	0.079***
	(0.006)	(0.006)	(0.008)	(0.008)	(0.007)	(0.008)
Good correction x Female partner		-0.030		0.053		-0.151*
		(0.060)		(0.080)		(0.088)
Bad correction x Female partner		0.241**		0.179		0.312**
		(0.096)		(0.127)		(0.143)
Any correction x Female partner	0.036		0.086		-0.044	
	(0.055)		(0.076)		(0.079)	
Partner's contribution x Female partner	-0.005	0.001	-0.007	-0.004	-0.002	0.007
	(0.007)	(0.008)	(0.010)	(0.010)	(0.010)	(0.011)
Good correction x High bias		-0.018		0.027		-0.076
		(0.071)		(0.093)		(0.106)
Bad correction x High bias		0.036		-0.069		0.173
		(0.100)		(0.129)		(0.146)
Any correction x High bias	0.019		0.011		0.017	
	(0.063)		(0.089)		(0.091)	
Female partner x High bias	-0.027	-0.015	-0.050	-0.039	-0.010	0.008
	(0.044)	(0.044)	(0.064)	(0.065)	(0.058)	(0.059)
Partner's contribution x High bias	-0.009	-0.008	-0.008	-0.010	-0.010	-0.006
	(0.008)	(0.009)	(0.012)	(0.012)	(0.010)	(0.011)
Good correction x Female partner x High bias		-0.003		-0.025		0.060
		(0.089)		(0.118)		(0.134)
Bad correction x Female partner x High bias		-0.177		-0.170		-0.257
		(0.137)		(0.186)		(0.198)
Any correction x Female partner x High bias	-0.048		-0.073		-0.011	
	(0.082)		(0.114)		(0.118)	
Partner's contribution x Female partner x High bias	0.007	0.003	0.011	0.008	0.005	-0.000
	(0.011)	(0.011)	(0.016)	(0.016)	(0.014)	(0.014)
Individual FE	✓	✓	✓	✓	✓	✓
Baseline mean	0.781	0.781	0.783	0.783	0.779	0.779
Baseline SD	0.414	0.414	0.413	0.413	0.415	0.415
Adj. R-squared	0.333	0.335	0.363	0.368	0.304	0.305
Observations	3173	3173	1670	1670	1503	1503
Individuals	463	463	244	244	219	219

Notes: This table presents the regression results of equation B1. Columns 1-2 includes all participants willingness to collaborate. Columns 3-4 present the corresponding results for women and columns 5-6 for men. Baseline mean and standard deviation are willingness to collaborate with partners who do not make any corrections. Standard errors in parentheses are clustered at the individual level. Significance levels: * 10%, ** 5%, and *** 1%.

Table C1: Response to corrections: The original contribution measure

Dependent variable:	Willing to collaborate (yes=1, no=0)									
Sample:	All					Female			Male	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Good correction	-0.208***	-0.238***		-0.272***	-0.269***		-0.304***	-0.197***		-0.230***
	(0.028)	(0.030)		(0.026)	(0.043)		(0.035)	(0.040)		(0.038)
Bad correction	-0.518***	-0.508***		-0.160***	-0.550***		-0.234***	-0.457***		-0.065
	(0.031)	(0.034)		(0.037)	(0.044)		(0.048)	(0.050)		(0.054)
Any correction			-0.267***			-0.313***			-0.213***	
			(0.024)			(0.033)			(0.033)	
Female partner	-0.003	-0.001	0.006	0.008	-0.009	0.001	0.004	0.007	0.012	0.012
	(0.016)	(0.017)	(0.014)	(0.014)	(0.021)	(0.019)	(0.019)	(0.026)	(0.021)	(0.022)
Partner's contribution			1.181***	1.192***		1.171***	1.164***		1.196***	1.234***
			(0.054)	(0.058)		(0.076)	(0.078)		(0.076)	(0.084)
Individual FE		✓	✓	✓	✓	✓	✓	✓	✓	✓
P-value: Good correction =Bad correction	0.000	0.000		0.013	0.000		0.252	0.000		0.014
Baseline mean	0.780	0.780	0.780	0.780	0.780	0.780	0.780	0.778	0.778	0.778
Baseline SD	0.414	0.414	0.414	0.414	0.414	0.414	0.414	0.416	0.416	0.416
Adj. R-squared	0.104	0.100	0.309	0.314	0.111	0.320	0.330	0.090	0.300	0.300
Observations	3180	3180	3180	3180	1670	1670	1670	1510	1510	1510
Individuals	464	464	464	464	244	244	244	220	220	220

Notes: This table reports the same estimation results as Table 3 but with the original contribution measure specified in the pre-analysis plan, and show that the results are robust to using the original measure. Baseline mean and standard deviation are willingness to collaborate with partners who do not make any corrections. Standard errors in parentheses are clustered at the individual level. Significance levels: * 10%, ** 5%, and *** 1%.

Table C2: Response to corrections of high vs. low ability people: The original contribution measure

Dependent variable:	Willing to collaborate (yes=1, no=0)					
Sample:	All		Female		Male	
	(1)	(2)	(3)	(4)	(5)	(6)
Good correction		-0.216***		-0.269***		-0.168***
		(0.033)		(0.044)		(0.047)
Bad correction		-0.178***		-0.291***		-0.055
		(0.051)		(0.068)		(0.069)
Any correction	-0.224***		-0.284***		-0.167***	
	(0.031)		(0.045)		(0.042)	
Female partner	0.006	0.007	0.001	0.002	0.011	0.010
	(0.014)	(0.014)	(0.019)	(0.018)	(0.021)	(0.021)
Partner's contribution	1.200***	1.196***	1.208***	1.196***	1.195***	1.222***
	(0.068)	(0.073)	(0.101)	(0.101)	(0.089)	(0.098)
Good correction x High ability		-0.137***		-0.079		-0.192**
		(0.052)		(0.070)		(0.076)
Bad correction x High ability		0.049		0.144		-0.051
		(0.072)		(0.094)		(0.108)
Any correction x High ability	-0.101**		-0.060		-0.130**	
	(0.047)		(0.066)		(0.065)	
Partner's contribution x High ability	-0.039	-0.004	-0.088	-0.070	0.025	0.064
	(0.112)	(0.118)	(0.152)	(0.154)	(0.167)	(0.183)
Individual FE	✓	✓	✓	✓	✓	✓
Baseline mean	0.780	0.780	0.783	0.783	0.778	0.778
Baseline SD	0.414	0.414	0.413	0.413	0.416	0.416
Adj. R-squared	0.310	0.315	0.320	0.331	0.301	0.303
Observations	3180	3180	1670	1670	1510	1510
Individuals	464	464	244	244	220	220

Notes: This table reports the same estimation results as Table 4 but with the original contribution measure specified in the pre-analysis plan, and show that the results are robust to using the original measure. Baseline mean and standard deviation are willingness to collaborate with partners who do not make any corrections. Standard errors in parentheses are clustered at the individual level. Significance levels: * 10%, ** 5%, and *** 1%.

Table C3: Response to corrections made by women vs. men: The original contribution measure

Dependent variable:	Willing to collaborate (yes=1, no=0)					
Sample:	All		Female		Male	
	(1)	(2)	(3)	(4)	(5)	(6)
Good correction		-0.268***		-0.338***		-0.180***
		(0.038)		(0.049)		(0.058)
Bad correction		-0.232***		-0.275***		-0.167**
		(0.054)		(0.069)		(0.084)
Any correction	-0.277***		-0.348***		-0.192***	
	(0.034)		(0.049)		(0.048)	
Female partner	-0.053	-0.072	-0.099	-0.090	-0.011	-0.063
	(0.049)	(0.052)	(0.070)	(0.074)	(0.069)	(0.072)
Partner's contribution	1.115***	1.109***	1.064***	1.070***	1.159***	1.147***
	(0.082)	(0.085)	(0.116)	(0.114)	(0.116)	(0.125)
Good correction x Female partner		-0.008		0.063		-0.090
		(0.046)		(0.061)		(0.071)
Bad correction x Female partner		0.143*		0.085		0.188*
		(0.077)		(0.108)		(0.105)
Any correction x Female partner	0.023		0.069		-0.035	
	(0.044)		(0.063)		(0.062)	
Partner's contribution x Female partner	0.124	0.163	0.201	0.182	0.061	0.168
	(0.107)	(0.113)	(0.152)	(0.159)	(0.150)	(0.157)
Individual FE	✓	✓	✓	✓	✓	✓
Baseline mean	0.780	0.780	0.783	0.783	0.778	0.778
Baseline SD	0.414	0.414	0.413	0.413	0.416	0.416
Adj. R-squared	0.309	0.314	0.321	0.331	0.299	0.301
Observations	3180	3180	1670	1670	1510	1510
Individuals	464	464	244	244	220	220

Notes: This table reports the same estimation results as Table 5 but with the original contribution measure specified in the pre-analysis plan, and show that the results are robust to using the original measure. Baseline mean and standard deviation are willingness to collaborate with partners who do not make any corrections. Standard errors in parentheses are clustered at the individual level. Significance levels: * 10%, ** 5%, and *** 1%.

Table C4: Response to corrections made by women vs. men: Heterogeneity by gender bias, the original contribution measure

Dependent variable:	Willing to collaborate (yes=1, no=0)					
Sample:	All		Female		Male	
	(1)	(2)	(3)	(4)	(5)	(6)
Good correction		-0.254***		-0.327***		-0.151*
		(0.051)		(0.063)		(0.083)
Bad correction		-0.211***		-0.173*		-0.268***
		(0.067)		(0.091)		(0.096)
Any correction	-0.275***		-0.318***		-0.213***	
	(0.045)		(0.059)		(0.072)	
Female partner	0.020	-0.004	-0.006	0.005	0.041	-0.029
	(0.061)	(0.065)	(0.086)	(0.092)	(0.086)	(0.087)
Partner's contribution	1.254***	1.247***	1.246***	1.275***	1.251***	1.203***
	(0.097)	(0.103)	(0.138)	(0.139)	(0.137)	(0.149)
Good correction x Female partner		-0.047		0.025		-0.159
		(0.066)		(0.087)		(0.102)
Bad correction x Female partner		0.161*		0.042		0.327**
		(0.094)		(0.130)		(0.129)
Any correction x Female partner	0.001		0.046		-0.064	
	(0.059)		(0.080)		(0.089)	
Partner's contribution x Female partner	-0.020	0.033	0.021	0.007	-0.047	0.100
	(0.130)	(0.140)	(0.186)	(0.202)	(0.182)	(0.186)
Good correction x High bias		-0.037		-0.037		-0.054
		(0.076)		(0.100)		(0.117)
Bad correction x High bias		-0.040		-0.188		0.175
		(0.104)		(0.132)		(0.159)
Any correction x High bias	-0.014		-0.074		0.035	
	(0.069)		(0.100)		(0.098)	
Female partner x High bias	-0.166*	-0.152	-0.198	-0.191	-0.135	-0.095
	(0.099)	(0.104)	(0.139)	(0.143)	(0.139)	(0.145)
Partner's contribution x High bias	-0.303*	-0.299*	-0.382*	-0.404*	-0.212	-0.148
	(0.164)	(0.167)	(0.226)	(0.217)	(0.236)	(0.254)
Good correction x Female partner x High bias		0.085		0.089		0.134
		(0.094)		(0.125)		(0.141)
Bad correction x Female partner x High bias		-0.026		0.068		-0.225
		(0.151)		(0.212)		(0.200)
Any correction x Female partner x High bias	0.053		0.061		0.067	
	(0.089)		(0.128)		(0.124)	
Partner's contribution x Female partner x High bias	0.322	0.289	0.382	0.354	0.272	0.186
	(0.214)	(0.225)	(0.301)	(0.311)	(0.303)	(0.318)
Individual FE	✓	✓	✓	✓	✓	✓
Baseline mean	0.781	0.781	0.783	0.783	0.779	0.779
Baseline SD	0.414	0.414	0.413	0.413	0.415	0.415
Adj. R-squared	0.310	0.315	0.322	0.332	0.298	0.299
Observations	3173	3173	1670	1670	1503	1503
Individuals	463	463	244	244	219	219

Notes: his table reports the same estimation results as Table B1 but with the original contribution measure specified in the pre-analysis plan, and show that the results are robust to using the original measure. Baseline mean and standard deviation are willingness to collaborate with partners who do not make any corrections. Standard errors in parentheses are clustered at the individual level. Significance levels: * 10%, ** 5%, and *** 1%.

Appendix D Experimental instructions: English translation

App: pt0

Page: Reg

Registration

Please fill out the following information in order for us to pay you after the session. Please make sure that they correspond to the information you registered on ORSEE.

N.B. Please capitalize only the first letter of your first name and last name.

Good examples: Marco Rossi; Maria Bianchi; Anna Maria Gallo

Bad examples: MARCO ROSSI; maria bianchi; Anna maria Gallo

- First name: [Textbox]
- Last name: [Textbox]
- Email address registered on ORSEE: [Textbox]

[Check if there are any same first names. If so, add an integer (starting from 2) at the end of the first name]

Page: Draw

Draw a coin

Please draw a virtual coin by clicking the button below.

[Draw]

[Assign random number ranging from 1 to 40]

Page: Wait

Your coin

You drew the following coin.

Please wait until the session starts.

Page: Excess

Please click an appropriate button

[I was chosen to participate] [I was chosen to leave]

Page: Intro

General instructions

Overview: This study will consist of **3 parts** and a follow-up survey and is expected to take **1 hour**. At the beginning of each part, you will receive specific instructions, followed by a set of understanding questions. You must answer these understanding questions correctly to proceed.

Your payment: For completing this study, you are guaranteed **2€** for your participation, but can earn up to **25€** depending on how good you are at the tasks. The tasks involve solving sliding puzzles, like the one shown below.

puzzle_2_0.png

Confidentiality: Other people participating in this study can see your first name. Aside from your first name, other participants will not see any information about you. **At the conclusion of the study, all identifying information will be removed and the data will be kept confidential**. If there is more than one participant with the same first name, we add a number at the end of your first name (e.g. Marco2).

General rules: During the study, please turn off your camera and microphone, and do not communicate with anyone other than us. Also, please do not reload the page or close your browser because it may make your puzzle unsolvable. If you have any questions or face any problems, please send us a private chat on Zoom.

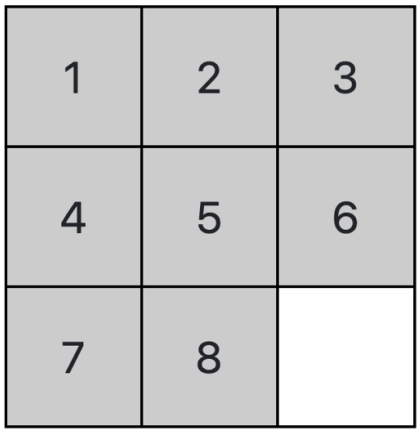
App: pt1

Page: Intro

Instructions for part 1 out of 3

In this part, you will solve the puzzle alone to familiarize yourself with it. You can solve as many puzzles as possible (but a maximum of 15 puzzles) in **4 minutes**. You will earn **0.2€ for each puzzle** you solve.

Your goal is to move the tiles and order them as follows:



puzzle_goal.png

Before you start, please go through the three examples below to understand how to solve the puzzle.

Example 1:

First, consider the following puzzle.

puzzle_1.png

You can only move the tiles next to an empty cell and the tile you choose is moved to the empty cell. So, in this puzzle, there are 3 moves you can make: move 3 down, move 5 right, and move 6 up.

Among the 3 moves, moving 6 up is the only correct move: by moving 6 up, you can solve the puzzle. The other moves do not solve the puzzle.

When you click a tile next to an empty cell, the tile will be moved to the empty cell. So, in this case, you should click 6 to move it up.

Example 2:

Next, consider the following puzzle.

puzzle_2_0.png

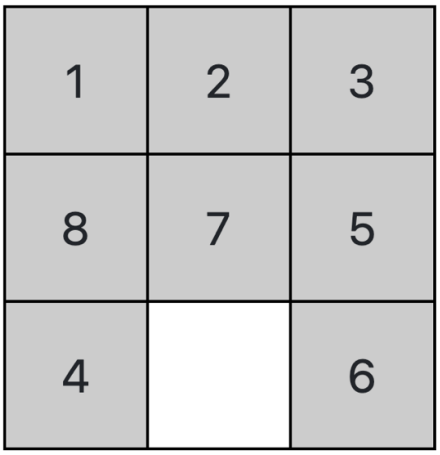
First, there are 2 moves you can make: move 2 right and move 3 up. Which moves should you make?

Observe that the only tiles that are not in the correct order are 3 and 6. So, you should move 3 up.

After moving 3 up, the puzzle will look like the one in example 1. Then you should move 6 up and the puzzle will be solved.

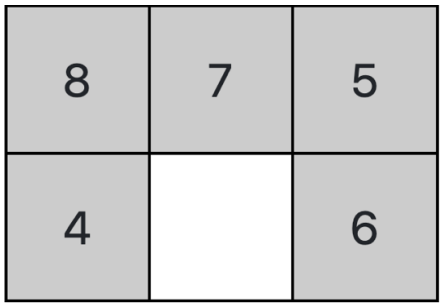
Example 3:

Finally, consider the following puzzle.



puzzle_3_0.png

This puzzle is a bit complicated but observe that the top row is already in the correct order. So, let's keep the top row as is, and think about the remaining part. **When the top row is in the correct order, you should always keep it as is**. So, think of this puzzle as the following simpler puzzle.



puzzle_3_0_2x3.png

You could solve the puzzle by trial and error. However, **after making the top row in the correct order, you should next make the left column in the correct order** to solve the puzzle faster. There are two moves you can make: move 4 right and move 7 down. Which is the faster way to make the left column in the correct order?

Let's try moving 4 right.

puzzle_3_1_bad_0.png

Now the only tile you can move is 8. So, let's move it down.

puzzle_3_1_bad_1.png

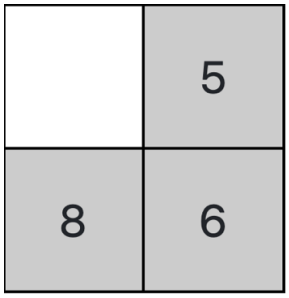
Now, if you ignore the top row which is already in the correct order, the only tile you can move is 7. So, let's move it to the left.

puzzle_3_1_bad_2.png

Then move 4 up, move 8 right, and move 7 down. Then you have made the left column in the correct order. You have moved tiles seven times until now.

puzzle_3_1_bad_3.png

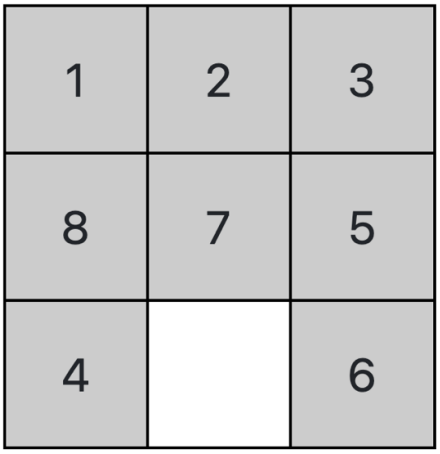
Now let's also keep the left column as is.



puzzle_3_1_bad_3_2x2.png

Then you can solve the puzzle by moving 5 left and then 6 up. With this method, **you have moved tiles nine times in total**.

Let's go back to the initial puzzle.

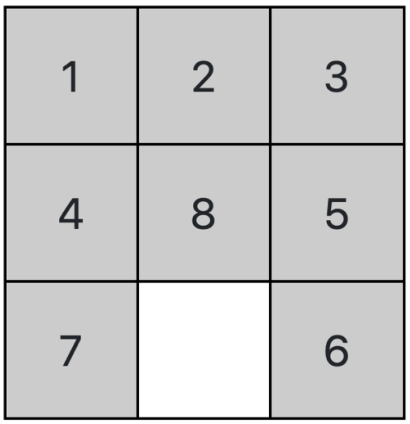


puzzle_3_0.png

This time, let's try moving 7 down.

puzzle_3_1_good.png

Then move 8 right, 4 up, and 7 left. Now you have made the left column in the correct order only with four moves.



puzzle_3_4_good.png

Let's keep the left column as is (as well as the top row).

puzzle_3_4_good_2x2.png

Now it's easy to solve the puzzle: move 8 down, 5 left, and 6 up. With this method, **you have only moved tiles seven times in total**.

Because there is a time limit, it's better to solve the puzzle with the minimum number of moves. **We call a move a good move if it makes a puzzle closer to the solution, and a bad move if it makes a puzzle far from the solution. There are no neutral moves: all moves are either good or bad.**

In summary: when you solve the puzzle, first make the top row in the correct order, then make the left column in the correct order. Always try to make the number of moves as small as possible.

<u>Understanding questions</u>:

Before you proceed, please answer the following understanding questions. After you answer, please click Next.

1. Which of the following statements is true?

- ✓In this part, I will work on the puzzles individually for 4 minutes and earn 0.2€ for each puzzle I solve.
- In this part, I will work on the puzzles in pairs for 4 minutes and earn 0.2€ for each puzzle we solve.
- In this part, I will work on the puzzles individually for 4 minutes, but I will not earn anything.

2. Which of the following puzzles is in the correct order?

- A
- ✓B

A

puzzle_2_0.png

B

puzzle_goal.png

3. What is the strategy you should use to solve the puzzle as fast as possible?

- First, make the left column in the correct order, then the bottom row. Always minimize the number of moves I make.
- First, make the top row in the correct order, then the right column. Always minimize the number of moves I make.
- ✔First, make the top row in the correct order, then the left column. Always minimize the number of moves I make.

4. Look at the following puzzle. Which is the good move?

- Move 4 down.
- ✔Move 7 left.

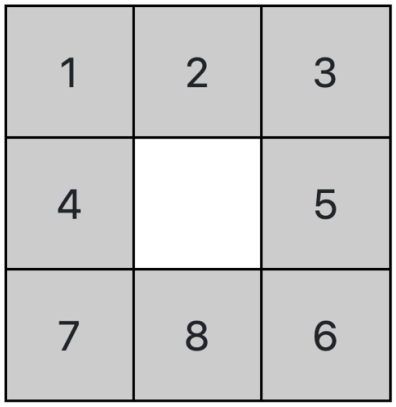
puzzle_3_3_good.png

5. Consider the puzzle in question 4. What is the minimum number of moves to solve the puzzle?

- 2
- 3
- ✔4

6. Look at the following puzzle. Which is the good move?

- ✓Move 5 left.
- Move 8 up.



puzzle_3_5_good.png

7. Consider the puzzle in question 6. What is the minimum number of moves to solve the puzzle?

- ✓2
- 3
- 4

Page: Ready

Be ready

[5 seconds time count]

Please be ready for the individual round.

Page: Game

Individual round

[4 minutes time count]

[max. 15 puzzles with increasing difficulty]

Page: Proceed

The individual round is over

The individual round is over. You have solved **xx puzzles**.

Please click Next to proceed.

App: pt2

Page: Intro

Instructions for part 2 out of 3

In this part, you will **choose your partner for part 3**, the next part.

Although you will not earn anything in this part, it is important to choose the best partner possible: in part 3, you will work on the puzzles for 12 minutes in a pair by moving the tiles in turn, and both you and your partner will earn 1€ for each puzzle you two solve. There is a maximum of 20 puzzles you and your partner can solve (so the maximum earning is 20€).

You will **meet 7 other people** participating in this session one by one and solve 1 puzzle together by moving tiles in turn as you would do in part 3. One of you will be randomly chosen to make the first move at the beginning of each puzzle. You will have a **2-minute limit** for each puzzle.

After solving the puzzle, you will **choose whether you want to work with this person in part 3 too**. This person or other people in this session will not see your choice. **You can choose as many people as you want**.

After you meet all the 7 people and state your choices, we will check all the choices you and the 7 other people have made, and decide each person's partner for part 3 as follows:

1. We randomly choose 1 person out of you and the other 7 people. Call this person Giovanni.
2. We then check if Giovanni has a "match": among people Giovanni has chosen, we check whether these people also have chosen Giovanni. If there is such a person, we make Giovanni and this person as partners for part 3.
3. If Giovanni has more than one match, we randomly choose one of the matches and make them as partners for part 3.
4. If Giovanni has not chosen anyone, the people Giovanni has chosen have not chosen Giovanni, or those people already have their partner, we put Giovanni on a waiting list and repeat points 1-3 above.
5. After we choose all people, we randomly match people on the waiting list as partners for part 3.

So, **even if you choose a particular person, you may not be able to work with that person in part 3**. So, choose everyone whom you want to work with in part 3.

<u>Understanding questions</u>:

Before you proceed, please answer the following understanding questions. After you answer, please click Next.

1. Which of the following statements is true?

- ✓In this part, I will choose my partner for part 3.
- In this part, I will work on the puzzles for 12 minutes in a pair by moving the tiles in turn.

2. How many people can you choose whom you want to work with in part 3?

- 1 person.
- 2 people.
- ✓As many people as you want.

3. Why is it important to choose the best partner for part 3?

- ✓ because how many puzzles I can solve in part 3 depends on my partner's moves.
- because my partner will solve puzzles for me.

4. Suppose you have chosen Giovanni and Valeria. However, while Valeria has chosen you, Giovanni has not. If we have randomly chosen you first, who will be your partner for part 3?

- Giovanni
- ✓Valeria
- Someone on the waiting list
- Randomly chosen from Giovanni and Valeria

5. Suppose you have chosen Giovanni and Valeria. However, unlike question 4, while Giovanni has chosen you, Valeria has not. If we have randomly chosen you first, who will be your partner for part 3?

- ✓Giovanni
- Valeria
- Someone on the waiting list
- Randomly chosen from Giovanni and Valeria

6. Suppose you have chosen Giovanni and Valeria. Also, both Giovanni and Valeria have chosen you. If we have randomly chosen you first, who will be your partner for part 3?

- Giovanni
- Valeria
- Someone on the waiting list
- ✓Randomly chosen from Giovanni and Valeria

7. Suppose you have chosen Giovanni and Valeria. Also, both Giovanni and Valeria have chosen you. However, we already matched Valeria with Giovanni before we choose you. Who will be your partner for part 3?

- Giovanni
- Valeria
- ✓Someone on the waiting list
- Randomly chosen from Giovanni and Valeria

8. Suppose you have not chosen anyone. Also, both Giovanni and Valeria have chosen you. If we have randomly chosen you first, who will be your partner for part 3?

- Giovanni
- Valeria

- ✓Someone on the waiting list
- Randomly chosen from Giovanni and Valeria

9. Suppose you have chosen Giovanni and Valeria. However, neither Giovanni nor Valeria has chosen you. If we have randomly chosen you first, who will be your partner for part 3?

- Giovanni
- Valeria
- ✓Someone on the waiting list
- Randomly chosen from Giovanni and Valeria

Page: Puzzle

Puzzle 1/2/3/4/5/6/7 out of 7

You are playing the puzzle with **[this person's ID]**

[2 minutes time count]

Page: Pref

Puzzle 1/2/3/4/5/6/7 out of 7

You have played the puzzle with **[this person's ID]**. Do you want to work with [this person's ID] in part 3?

[Yes, No]

App: pt3

Page: Partner

Your partner for part 3

Based on your and the 7 other people's choices, **[the partner's ID]** became your partner for part 3.

Page: Intro

Instructions for part 3 out of 3

In this part, you will work on the puzzles with your partner for **12 minutes** by moving the tiles in turn, and both you and your partner will earn **1€ for each puzzle** you two solve. There is a maximum of 20 puzzles you and your partner can solve (so the maximum earning is 20€). As in part 2, one of you will be randomly chosen to make the first move at the beginning of each puzzle.

Understanding questions:

Before you proceed, please answer the following understanding questions. After you answer, please click Next.

1. Which of the following statements is true?

- ✓In this part, you and your partner will both earn 1€ for each puzzle you two solve, which means you will earn 1€ for each puzzle you two solve.
- In this part, you and your partner will earn 1€ for each puzzle you two solve, which means you will earn 0.5€ for each puzzle you two solve.

2. You and your partner…

- ✓will work on the puzzles for 12 minutes by moving the tiles in turn. Which of you will make the first move is randomly determined at the beginning of each puzzle.
- will work on the puzzles for 12 minutes. Which of you will make the first move is randomly determined at the beginning of this part and fixed afterward.

Page: Ready

Be ready

[5 seconds time count]

Please be ready for the group round.

Page: Game

Puzzle 1/2/3/…/20

Your partner: **[the partner's ID]**

[12 minutes time count]

[max. 20 puzzles with increasing difficulty]

Page: Proceed

The group round is over

The group round is over. You have solved **xx puzzles**.

Please click Next to proceed.

<u>App: pt4</u>

Page: Intro

A follow-up survey

As the last task, we will ask you a series of questions in which there are no right or wrong answers. We are only interested in your personal opinions. We are interested in what

characteristics are associated with people's behaviors in this study. **The answers you provide will in no way affect your earnings in this study and are kept confidential.**

Please click Next to start the survey.

Page: SurveyASI

Survey page 1 out of 2

Below is a series of statements concerning men and women and their relationships in contemporary society. Please indicate the degree to which you agree or disagree with each statement.

- Women are too easily offended.
- Many women are actually seeking special favors, such as hiring policies that favor them over men, under the guise of asking for "equality."
- Men should be willing to sacrifice their own wellbeing in order to provide financially for the women in their lives.
- Many women have a quality of purity that few men possess.
- No matter how accomplished he is, a man is not truly complete as a person unless he has the love of a woman.
- Women exaggerate problems they have at work.

[Choices: Strongly agree, Agree a little, Neither agree nor disagree, Disagree a little, Strongly disagree]

Page: SurveyDem

Survey page 2 out of 2

Please tell us about yourself and your opinion about this study.

- Your age: [Integer]
- Gender: [Male, Female]
- Region of origin: [Northwest, Northeast, Center, South, Islands, Abroad]
- Field of study: [Humanities, Law, Social Sciences, Natural Sciences/Mathematics, Medicine, Engineering]
- Degree program: [Bachelor, Master/Post-bachelor, Bachelor-master combined (1st, 2nd, or 3rd year), Bachelor-master combined (4th year or beyond), Doctor]
- What do you think this study was about? [Textbox]
- Was there anything unclear or confusing about this study? [Textbox]
- Were the puzzles difficult? [Difficult, Somewhat difficult, Just right, Somewhat easy, Easy]
- Do you have any other comments? (optional) [Textbox]

Page: ThankYou

Thank you for your participation

Thank you for your participation. You have completed the study.

Your earnings:

- **2€** for your participation.
- **xx.x€** for the puzzles you solved in part 1.
- **xx€** for the puzzles you and your partner solved in part 3.

Thus, you have earned **xx.x€** in this study. We will pay you your earnings via PayPal within 2 weeks. If you haven't received your earnings after 2 weeks, please contact us.

Optional: If you would like to know the results of this study, we are more than happy to send you the working paper via email once we finish this study.

[No, I do not want to receive the working paper] [Yes, I want to receive the working paper]

App: pt99

Page: ThankYou

Thank you for showing up

Thank you for showing up in this study. You will receive the show up fee of **2€** via PayPal within 2 weeks. If you haven't received your earnings after 2 weeks, please contact us.