Gender Differences in the Cost of Corrections in Group Work

Yuki Takahashi*

Click here for the latest version

December 7, 2022

Abstract

Collaboration is an integral component of workplace environments, but it also involves correcting one's colleagues. Using a quasi-laboratory experiment, I study whether people dislike collaborating with someone who corrects them and whether men dislike women's correction more. I find that people are less willing to collaborate with others who have corrected them, even if the correction improved group performance. High-ability people also respond negatively to corrections, suggesting that the negative response is unlikely due to their misunderstanding of correction quality. In addition, although only suggestive, men respond more negatively to women's performance-improving corrections but not to women's performance-deteriorating corrections, while women respond roughly equally negatively to any corrections by any gender. Their beliefs about gender differences in abilities cannot explain these differential responses. These findings suggest that a behavioral bias distorts the selection of talent and penalizes those who correct others' mistakes, especially when women correct men.

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

*Department of Economics, European University Institute. Via dei Roccettini 9, 50014 Fiesole, Italy. Email: yuki.takahashi@eui.eu. 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, Boon Han Koh, Annalisa Loviglio, Valeria Maggian, Natalia Montinari, Vincenzo Scrutinio, Hans Sievertsen, and many others, including participants at the CSQIEP Job Market Seminar, Stanford Institute for Theoretical Economics conference, Warwick Economics PhD Conference, Webinar in Gender and Family Economics, seminars at Ca' Foscari University, NHH, Osaka University, the University of Amsterdam, and the University of Bologna. Tommaso Batistoni, Philipp Chapkovski, Christian König genannt Kersting, and the 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 write the experimental instructions in Italian. Michela Boldrini and Boon Han Koh conducted quasi-laboratory experiments before me and kindly answered my questions about their implementations. Lorenzo Golinelli provided excellent technical and administrative assistance and John Acker provided excellent copyediting service. 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). The pre-abalysis plan is also in Online Appendix D and explanation of deviations from the plan is in Online Appendix A.

1 Introduction

Collaboration is a core element of workplace productivity, as most workplaces require group work (Jones 2021; Lazear and Shaw 2007; Wuchty, Jones, and Uzzi 2007). These interactions often involve correcting one's colleagues. For example, a worker may correct graphs with wrong numbers in the presentation slides one of their colleagues has prepared, or a seminar audience may point out an error in the identification assumption that the presenter is making. These corrections are essential for the groups to function well, but can also damage the collaborative relationship if people dislike being corrected. This potential interpersonal friction can be detrimental to group efficiency by its own merit but also through deterioration of workplace climate because workplace climate is an important determinant of productivity (Alan, Corekcioglu, and Sutter 2022; Edmans 2011; Guiso, Sapienza, and Zingales 2015; Haeckl and Rege 2022). Further, women may experience stronger interpersonal friction with men because some men dislike to be led by women (Abel 2022; Born, Ranehill, and Sandberg 2022; Chakraborty and Serra 2022; Husain, Matsa, and Miller 2021). This practice can contribute to gender gaps in labor market outcomes (Blau and Kahn 2017).

This paper studies whether people dislike collaborating with someone who corrects them and whether men dislike woman's correction more. I define collaboration as working with others toward the same goal, and correction as overriding what others do. Answering these questions using observational 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. In the experiment, participants are allocated to groups of eight people and are paired with each of the other group members sequentially, in random order, to solve a collaborative task together. Each time participants finish the task, they state whether they would like to collaborate with their current partner for the same task in the final stage of the experiment. This final stage is the main source of earnings for participants and thus provides a strong incentive for them to select as good a collaborator as possible. For the collaborative task, I use Isaksson (2018)'s number-sliding puzzle, which allows us to calculate an objective measure of each participant's contribution to the collaborative task and to classify each move as good (moving the puzzle closer to the solution) or bad (moving the puzzle further away from the solution). The puzzle also allows us to define a correction – reversing a partner's move – to make it comparable across different participants and to objectively classify it as either good or bad. Participants are informed at the beginning of the experiment of the notion of good and bad moves, how to solve the puzzles efficiently, and how the collaborator will be selected for the final stage.

I find that participants understand the notion of good and bad moves; the more a participant contribute to solving the puzzle, the more likely they are asked to be a collaborator. This is in line with what one would expect, and validates my experimental design. Nonetheless, after controlling for the individual contributions, participants are less willing to collaborate with someone who has corrected their moves, even if the corrections moved the puzzle closer to the solution. This is not because participants misunderstood good corrections as bad ones, as even high ability participants

– those who should be better able to identify good and bad corrections – respond negatively to corrections. Thus, the negative response is likely to be irrational. Although only suggestive (at the 10% statistical significance level), I also find evidence that men respond more negatively to women's good corrections but not to women's bad corrections: men dislike women correcting their mistakes but did not dislike women making mistakes on their own. On the other hand, women respond equally negatively to both women's and men's good and bad corrections. These findings are unlikely to be due to women's or men's beliefs about the differences in women's and men's abilities in solving the puzzle: women and men contribute equally well to the puzzle, and neither women nor men under- or overestimate women's contribution. Taken together, these findings suggest that a behavioral bias distorts the optimal selection of talents and penalizes those who correct others' mistakes, and that men may exhibit stronger bias when women correct them.

Isaksson (2018) uses the same puzzle and shows that women underclaim their contribution, especially in difficult puzzles, and that men correct their partners more often than women. Using Isaksson's puzzle and building on their findings, I answer different research questions with a different experimental design. Specifically, I examine whether receiving corrections reduces one's willingness to collaborate with that person and whether men react more negatively to women's corrections using a design adapted from Fisman et al. (2006, 2008)'s speed dating experiments but for working partner preference instead of romantic partner preference.

This paper's contribution is twofold. First, it contributes to the literature on workplace climate and productivity by showing that interpersonal frictions can distort group efficiency, and that the frictions can have a stronger effect on women. My findings complement Alan, Corekcioglu, and Sutter (2022), who find that a better workplace climate increases worker satisfaction and the degree of mutual reciprocation, while reducing toxic competition and worker turnover. Alan et al. argue that improved manager-worker relationships are the most likely mechanism. Relatedly, Haeckl and Rege (2022) find that supportive leaders increase worker satisfaction and engagement. Aside from Alan et al. and Haeckl and Rege, my findings also relate to the organizational economics literature: the literature finds that firms with high employee satisfaction exhibit higher stock prices (Edmans 2011) and that a firm perform better when its workers perceive their managers as trustworthy and ethical (Guiso, Sapienza, and Zingales 2015). In addition, I show that the same environment can affect women and men differently, which corroborates Dupas et al. (2021), who find female economists receive more patronizing and hostile questions during seminars, and Folke and Rickne (2022), who find that women in male-dominant jobs receive more harassment – as do men in female-dominant jobs.

Second, this paper contributes to the literature on differential treatment of women's opinions by showing that women's corrections can receive stronger negative reactions. My findings primarily complement Guo and Recalde (2022), who find that group members correct women's ideas more often than men's, and Coffman, Flikkema, and Shurchkov (2021), who find that group members are less likely to choose women's answers as a group answer in male-typed questions.

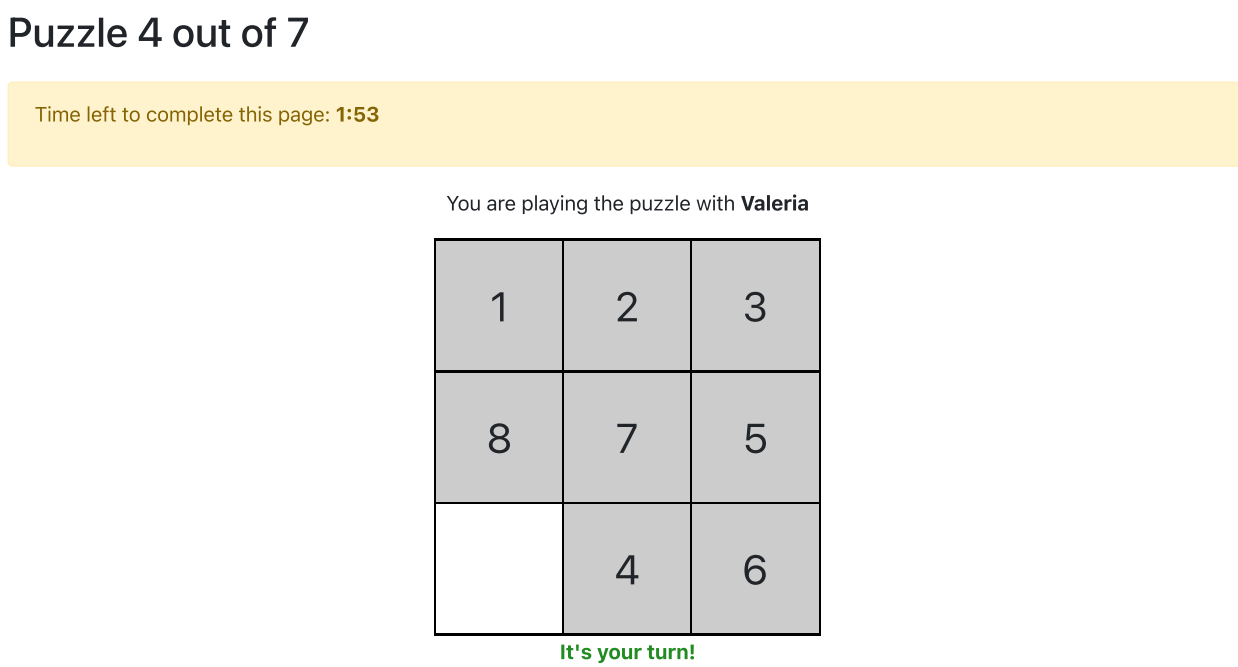
The remainder of the paper proceeds as follows. In section 2, I describe the experimental design,

procedure, and implementation. Next, I describe the data obtained from the experiment in section 3. Then, I provide a simple theoretical framework to show how a rational agent would behave in section 4. Afterward, I proceed to empirical analysis: I present the empirical strategy in section 5 and present the results in 6. I show the robustness of the results in section 7. Finally, I conclude the paper in section 8.

2 Experiment

Introducing the quasi-laboratory format I run the experiment in a quasi-laboratory format where we experimenters and the participants are connected via Zoom throughout the experiment, but turn off participants' cameras and microphones except at the beginning of the experiment. Aside from that participants participate remotely using their computers, the experiment is conducted as it would be in a physical laboratory. Appendix B 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 in the 4th round of the puzzle and makes their move. All the texts are in Italian in the experiment.

The collaborative task For the collaborative 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.[1] This puzzle has nice mathematical properties: I can define the puzzle's difficulty and classify a

1. Each participant has to make a move during their turn; they cannot pass.

given move as either good or bad via the Breadth-First Search algorithm.[2] Based on the number of good and bad moves a participant makes, I can calculate individual contributions to the task: the contributions are measured by net good moves, the number of good moves minus the number of bad moves an individual makes in a given puzzle.

The puzzle also allows me to objectively compare the quality of corrections by different participants.[3] Further, puzzle-solving captures an essential characteristic of collaborative work 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). This partial observability allows participants to engage in motivated reasoning (Kunda 1990; Chance and Norton 2015; Gino, Norton, and Weber 2016), interpreting the quality of corrections in a self-serving manner.

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.[4] Since a correction is also a move, it is also either good or bad. 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 ensure they understand the instructions.

2.1 Design and procedure

Registration

Upon receiving an invitation email to the experiment, participants register for the session they want to participate in, and upload their ID documents as well as a signed consent form.[5]

Pre-experiment

Participants enter the Zoom waiting room on the day and at the time of the session they have registered for.[6] They receive a link to the virtual room for the experiment and enter their first name, last name, and the email address they used in the registration. They also draw a virtual coin numbered from 1 to 40 without replacement.

As participants arrive and are verified, I admit them to the Zoom meeting room one by one and rename them using the first name they have just entered. This information is necessary to match up

2. The difficulty is defined as the number of moves away from the solution; a good move is defined as a move that reduces the distance (in number of moves) to the solution, while a bad move is defined as a move that increases the distance to the solution.

3. Because some corrections happen early in the puzzle and the others later in the puzzle, Thus, what I capture in the analysis is the average effect of a correction.

4. This assumes that both players are trying to solve the puzzle; I show in section 7 that the results are robust to the exclusion of puzzles where either player might not be trying to solve the puzzle.

5. I recruited a few more participants than needed for each session in case some participants did not show up to the session.

6. The Zoom link is sent with an invitation email; I checked that each participant in the waiting room indeed had registered for that session before admitting them to the main room.

their earnings in this experiment to 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 a 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 asking them to respond verbally via their 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 drawn number to leave.[7] Those who leave the session receive a 2€ show-up fee. Figure 2 shows the 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 giving the rules of the experiment and take questions on Zoom. Once participants start the main part of the experiment, they can only communicate with the experimenter via Zoom's private chat.

Part 1: Individual practice stage

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

At the beginning of part 1, I explain participants in depth how to solve the puzzle efficiently (in minimum moves) and provide comprehension questions about the solution strategies; see the instructions in Appendix C.[8]

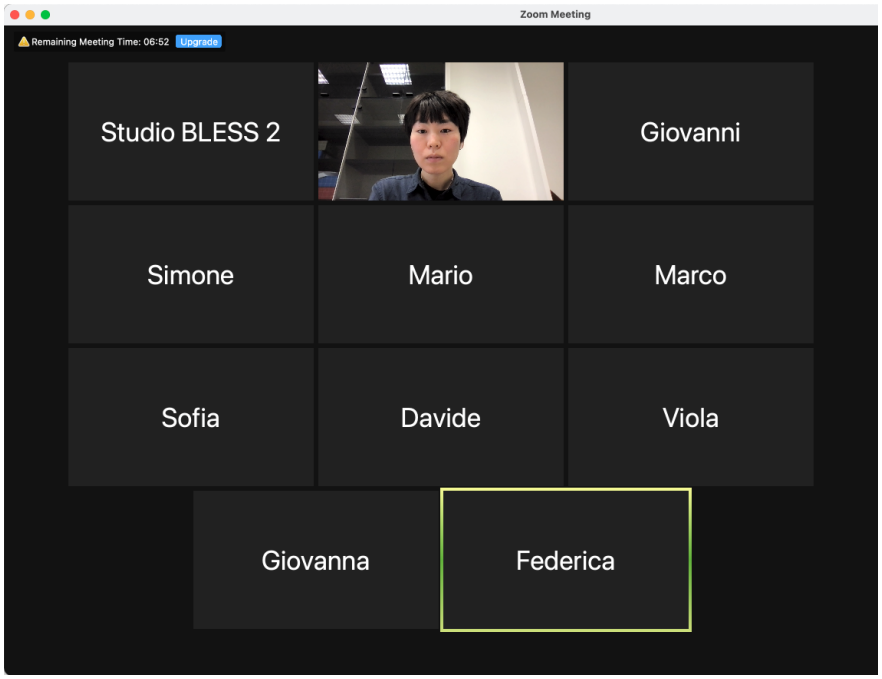
Part 2: Collaborator selection stage

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 divided into a group of 8, with participants of similar ability measured in part 1 placed in the same group. This is to reduce ability differences among participants, and participants are not told about this grouping criterion.

7. 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 the next number until the number of participants is 16. I share my computer screen during this process so that participants see the numbers are actually drawn randomly.

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

Figure 2: Zoom screen



Notes: This figure shows the 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.

Second, participants are paired with another randomly chosen participant in the same group, and they 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 solve the puzzle within 2 minutes, they finish the puzzle without solving it. Participants are allowed to reverse – that is, correct – their partner's move.[9] Each participant's contribution to a given puzzle is measured by net good moves. Figure 1 shows a sample puzzle screen where one participant is paired with another participant called Valeria and is making their move. Each partner's first name is displayed on the computer screen throughout the puzzle, and when participants select their collaborator, to subtly inform the partner'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 (so each participant solves the puzzle with seven different partners). In each round, participants solve another puzzle in pairs, 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 all seven rounds. I set

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

the minimum number of moves to solve the puzzles to be 8 based on a pilot so that the puzzles are neither too easy nor too difficult to solve.

At the end of part 2, participants are paired according to the following algorithm, adapted from Fisman et al. (2006, 2008):

1. For every participant 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 participants that have already 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.

At the beginning of part 2, I explain in depth this pairing algorithm along with comprehension questions so that the collaborator preference statement is incentivized.

The assumption for this matching algorithm to be incentive compatible is that payoff is the primary concern for participants. While this may not be a valid assumption in the real world because people may also care about their ability relative to their collaborator, for example, these factors do not play important roles in my experiment. Other individual-specific factors are controlled for by exploiting within-subject variation. Abilities of the past and the future partners' abilities are random because of the random pairing of participants within the same group.

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, which is informed to both participants as in part 2. They can solve as many puzzles as possible (maximum 20), with increasing difficulty.

Post-experiment

Each participant answers a short questionnaire, which consists of (i) the six hostile and benevolent sexism questions used by Stoddard, Karpowitz, and Preece (2020) with US college students and (ii) their basic demographic information and their impressions about the experiment.[10] The answer to their demographic information is used to know participants' characteristics as well as casually check whether they have anticipated that the experiment was about gender, for which I did not find any evidence.

10. I initially planned to use a gender bias measure, constructed from the hostile and benevolent sexism questions, to test whether those with higher gender bias responded more negatively to women's corrections. However, I could not have enough variation in this gender bias measure, so decided not to report it in the main text. See Online Appendix A for more detail.

After participants answer all the questions, I tell them their earnings and let them leave the virtual room and close Zoom. They later 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), all registered in 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).[11] The first two conditions were to reduce noise coming from differences in socio-demographic backgrounds and race or/and ethnicity that may be inferred from participants' first names or/and voices, and the last condition was to reduce experimenter demand effects.[12] The number of participants was determined by a power simulation in the pre-analysis plan to achieve 80% power.[13] The experiment was pre-registered with the OSF and the pre-analysis plan is in Online Appendix D.[14] Online Appendix A explains deviations from the pre-analysis plan.

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 a maximum of 25€ and a minimum of 2€, including the 2€ show-up fee. Table 1 describes the participants' characteristics. The table shows that 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 while 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).[15] Also, most female and male participants are either bachelor's or master's students (97% female and 94% male), and only a few are PhD students.[16]

3 Data description

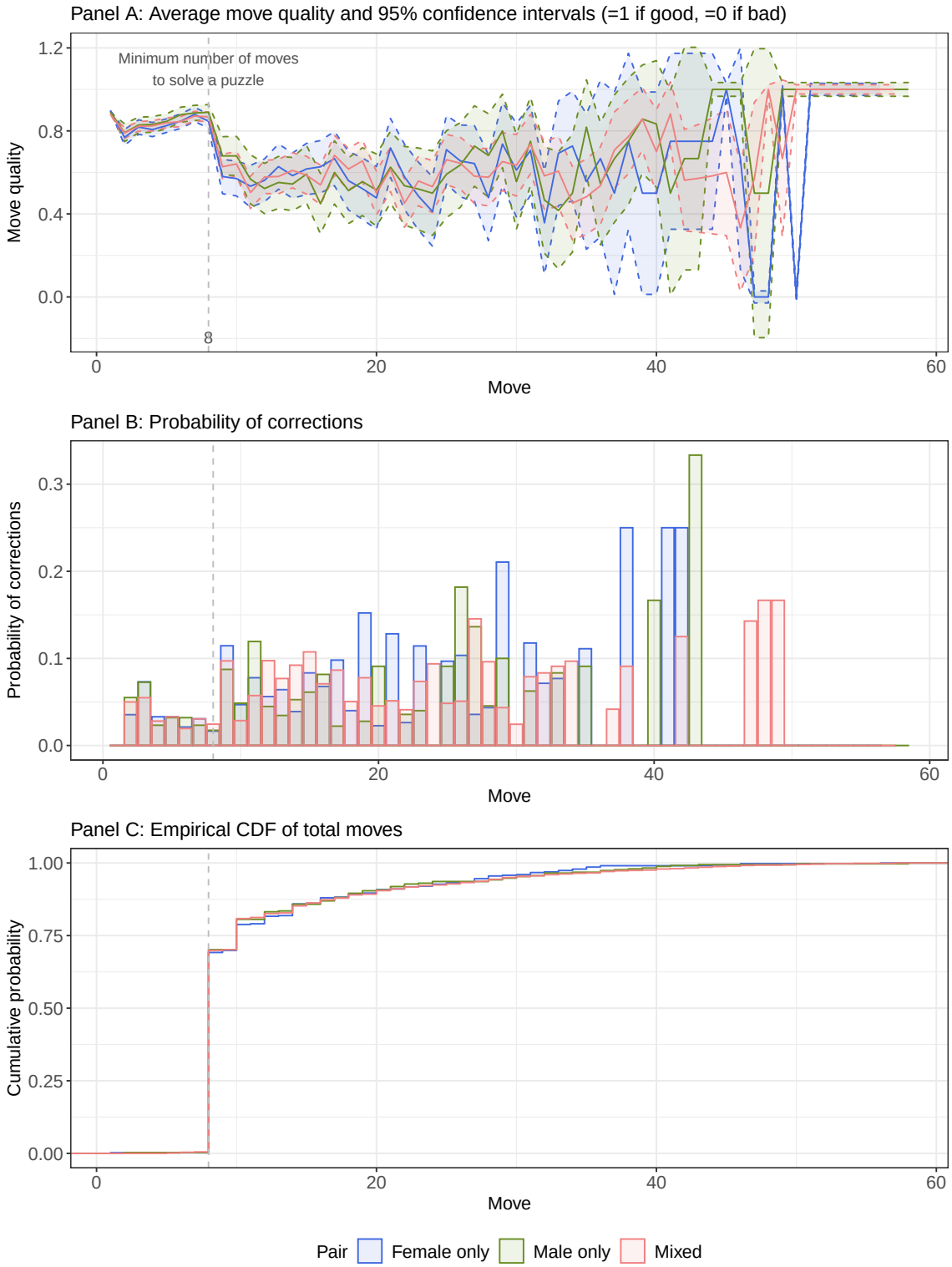
I use part 2 data in the analysis, as that is where we can observe collaborator selection decisions. I aggregate the move-level data for each puzzle so that we can associate behaviors with the puzzle to the collaborator selection decisions.

3.1 Move-level data

Figure 3 shows average move quality across moves along with 95% confidence bands (Panel A), probability that a correction is happening in a given move (Panel B), and empirical CDF of total

11. The laboratory prohibits deception, so no participant had participated in an experiment with deception.
12. Despite only recruiting Italy-born people, 1 male participant answered in the post-questionnaire that he was from abroad. I included this participant in the analysis anyway but the results are robust to excluding him.
13. 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 the exclusion of these 16 participants.
14. You can find the OSF registry at the following URL: https://osf.io/tgyc5.
15. Individual fixed effects in the analysis control for participants' major. However, I do not run heterogeneity analysis by major because one's major choice is endogenous to one's gender.
16. No economics PhD students participated in the experiment.

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



Notes: The average move quality along with 95% confidence intervals (panel A), the probability of corrections in each move (panel B), and the empirical CDF of total moves (panel C) separately for females only (blue), males only (green), and mixed gender pairs (red). 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 on the 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.

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 (within Italy)								
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
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.

moves (Panel C) for female-only pairs (blue), male-only pairs (green), and mixed-gender pairs (red). Panel A shows no statistically significant differences in move quality by one's own gender or the gender of one's partner. Panel B shows that corrections happen across the moves, but there are no systematic differences in the probability that correction is happening by one's own gender or the gender of one's partner. Panel C shows that about 70% of the puzzles are solved within a minimum number of moves (the minimum number of moves is 8) and shows that one's own gender or the gender of one's partner does not matter in how fast participants solve the puzzle.

3.2 Puzzle-level data

Table 2 describes own (panel A) and partner's puzzle behaviors (panel B) and puzzle outcomes (panel C). Panel A shows no gender differences in puzzle-solving ability: for both the contributions in part 2 and the number of puzzles solved in part 1, the difference between female and male participants is statistically insignificant at 5% and quantitatively insignificant.[17,18] This is consistent with Isaksson (2018), who also finds no gender difference in contribution or number of puzzles solved alone using the same puzzle, suggesting that any gender differences I find are unlikely to come from

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

18. 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 the individual level).

the ability differences between female and male participants. Panel A also shows that there are no gender differences in propensity to correct partners, unlike Isaksson (2018), who finds that men correct their partners more often than women, although that result is from move-level data. Finally, the last row of Panel A shows that male participants are slightly more likely to have female partners, although only three percentage points more.

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

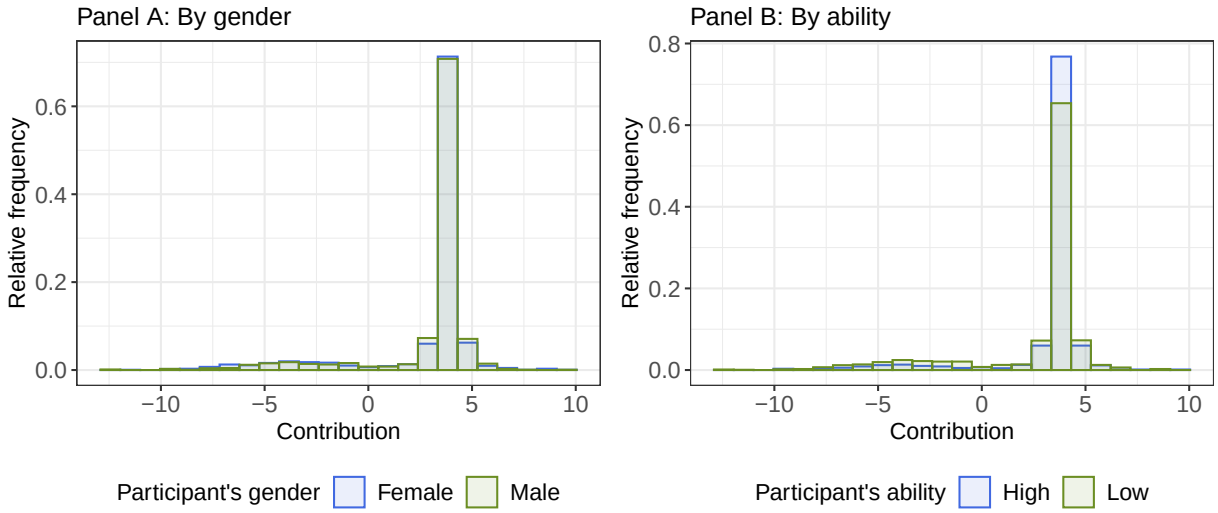
	Female (N=1708)		Male (N=1540)		Difference (Female – Male)		
	Mean	SD	Mean	SD	Mean	SE	P-value
Panel A: Own behaviors							
Contribution	2.98	2.93	3.14	2.64	-0.16	0.10	0.11
# puzzles solved in part 1	8.36	2.41	8.80	2.34	-0.44	0.22	0.05
Any correction	0.15	0.36	0.16	0.36	0.00	0.01	0.85
Good correction	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
(Fraction of female partners)	0.51	0.50	0.54	0.50	-0.03	0.02	0.03
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).

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

Panel B shows that the puzzle-solving ability, as well as the propensity to correct partners' moves (both mistakes and right moves) was the same for partners paired with female and male participants, suggesting that the 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 partners in 15-16% of the total puzzles, of which 12-13% are good corrections, and

Figure 4: Distribution of contributions



Notes: This figure shows the distribution of individual contributions 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 contributions, and panel B further shows that among high-ability participants, a 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).

5-6% are bad corrections, and there are no gender differences in the propensity to be corrected.[19]

Panel C shows that participants state they want to collaborate with the paired 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 correct their partner's moves consecutively in 4% of the puzzles.[20] There is no gender difference in any of these outcomes, suggesting any gender differences cannot be attributed to the imbalance in these outcomes.[21]

3.3 Across-round balance

Figure 5 plots the 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 (relative to round 1), separately for female (blue) and male participants (green).

First, there is some unbalance in partner's gender across rounds between female and male participants (Panel A), with female (male) participants more (less) likely to be paired with a female

19. The percentage of good corrections and bad corrections do not sum up to the percentage of all corrections because 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.

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

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

partner in round 1, but the difference is not statistically significant for rounds 2-7.

Second, there are no systematic gender differences in puzzle outcomes across rounds (Panels B-H), suggesting that female and male participants behave similarly across rounds. One difference could be good and bad corrections, with female participants making slightly more bad corrections and slightly fewer good corrections. However, as shown in Table 2, these differences are statistically insignificant.

Last, we see that in rounds 6 and 7, participants are less willing to collaborate, experience more corrections, and are less likely to solve the puzzle. Although these are all outcomes of a particular pair that is randomly formed, they can simply be correlations. Still, one may wonder whether rounds 6 and 7 are driving the results. I will show in section 7 that the results are robust to the exclusion of these rounds.

4 Theoretical framework

In this section, I present a simple theoretical framework to provide a rational agent's benchmark behaviors.

I consider a rational agent i who maximizes their expected payoff in a given round t by deciding whether they are willing to collaborate with a potential collaborator j with whom they have just played one puzzle, conditional on the history of decisions i has made about other potential collaborators with whom they have played the puzzle up to the current round t and with whom they will play the puzzle in the future rounds. Since with whom to be paired in which order is randomized, I simply denote the history and the future by t, consider them as exogenous, and normalize the payoff of not being willing to collaborate with j as 0 for each round t.

The payoff is increasing with 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 with j's ability as perceived by i.

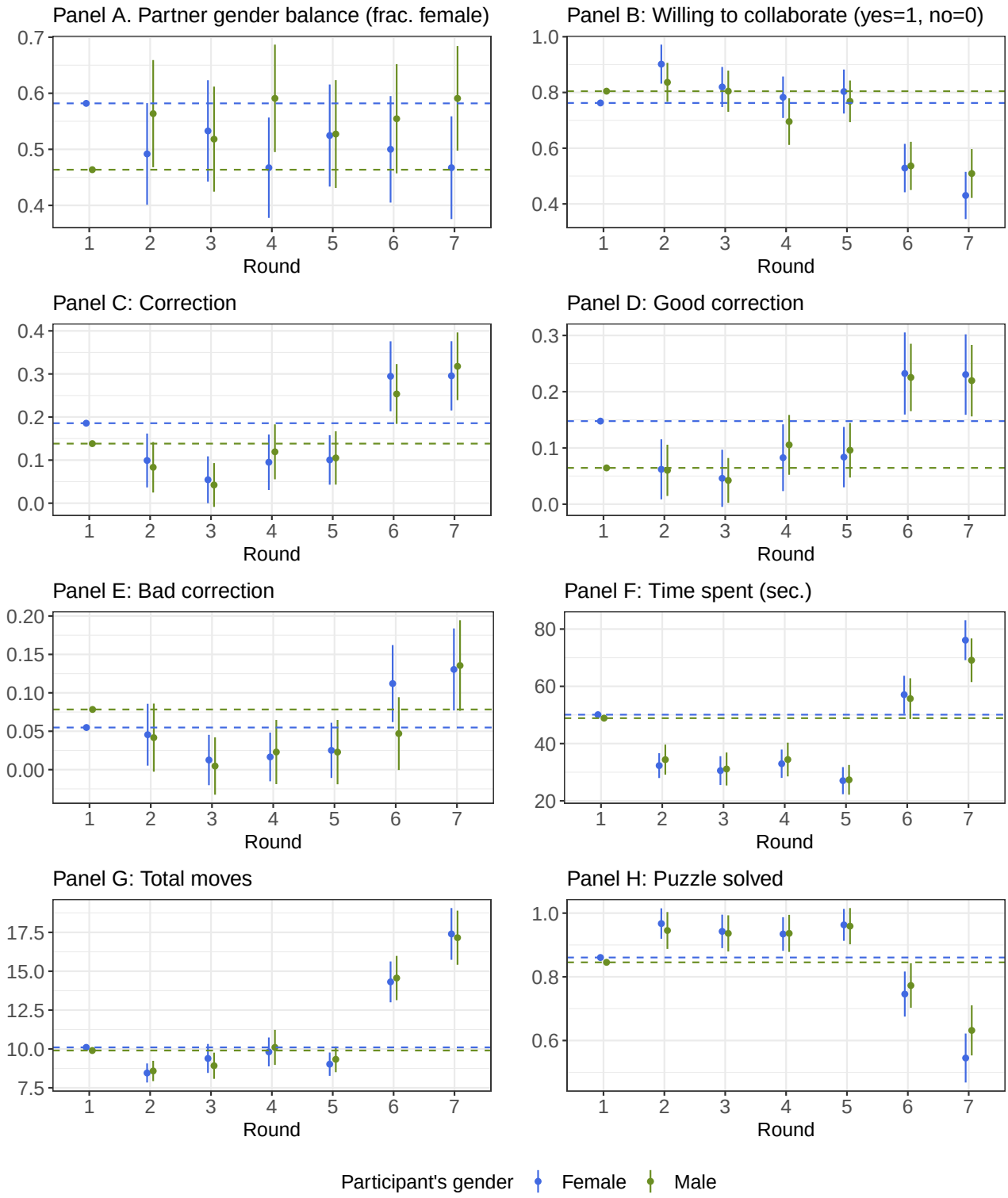
Thus, i would face the following problem:

$$\max_{Accept \in \{0,1\}} \mathbb{1}[Accept = 1] \times E_{\mu_j}[\pi_t(\mu_j(\tilde{a}_j, c_j^q, f_j))|\theta, \omega, t], \quad \partial\pi_t/\partial\mu_j > 0,\ \partial\mu_j/\partial\tilde{a}_j > 0 \tag{1}$$

where each term is defined as follows:

- $Accept$: whether i is willing to collaborate with j (=0 if no, =1 if yes)
- μ_j: i's belief about j's ability
- $\tilde{a}_j$: j's ability perceived by i
- c_j^q: j's correction (=1 if j corrected i, =0 if j did not correct i), which is either good ($q = g$) or bad ($q = b$).
- f_j: j's gender (=1 if female, =0 if male)
- θ: i's belief about their ability relative to other participants in the session (>0 if higher, =0 if same, <0 if lower)
- ω: j's belief about women's ability relative to men (>0 if higher, =0 if same, <0 if lower)

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 the 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.

where $\mathbb{1}$ is an indicator function. Although θ and ω could depend on t, I omit the dependence on t for simplicity because t is exogenous.

If i can fully observe j's move quality and i is fully rational, then c_j^q $(q = g, b)$ and f_j do not convey any information about j's ability and are irrelevant for i's decision making. This is true regardless of whether the correction is good or bad. However, since i can only partially observe j's move quality, j's corrections and gender convey information about j's ability, even if i is fully rational.[22]

First, keeping j's ability as perceived by i fixed, the information j's correction conveys depends on θ. 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 a less negative/more positive signal than a bad correction regardless of θ. Thus, we have the following proposition:

Proposition 1. *A rational agent i is less willing to collaborate with j when j made a bad correction than when j made a good correction, regardless of i's belief about their own ability. That is:*

$$\partial\mu_j/\partial c_j^b < \partial\mu_j/\partial c_j^g \ \forall\theta \tag{2}$$

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

Proposition 2. *A rational agent i with high puzzle-solving ability is more willing to collaborate with j when j made a good correction and is less willing to collaborate with j when j made a bad correction, compared to another rational agent with low puzzle-solving ability. This is true regardless of their belief about their own ability. That is:*

$$\begin{aligned} \partial\mu_j/\partial c_j^g|_{i\text{'s ability is high}} &> \partial\mu_j/\partial c_j^g|_{i\text{'s ability is low}} \ \forall\theta \\ \partial\mu_j/\partial c_j^b|_{i\text{'s ability is high}} &< \partial\mu_j/\partial c_j^b|_{i\text{'s ability is low}} \ \forall\theta \end{aligned} \tag{3}$$

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

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

Proposition 3. *A rational agent i's willingness to collaborate with j when j is a woman and made a correction, relative to when j is a man and make a correction, depends on i's belief about women's ability relative to men's. This is true regardless of i's belief about their own ability and holds for both good and bad corrections. That is:*

$$\begin{aligned} \partial^2 \mu_j / \partial c_j^q \partial f_j > 0 \; \forall \theta, q \;\; \textit{if} \; \omega > 0 \\ \partial^2 \mu_j / \partial c_j^q \partial f_j < 0 \; \forall \theta, q \;\; \textit{if} \; \omega < 0 \end{aligned} \tag{4}$$

In particular, if they believe women and men have the same ability, then j's gender does not matter. That is:

$$\partial^2 \mu_j / \partial c_j^q \partial f_j = 0 \; \forall \theta, q \;\; \textit{if} \; \omega = 0 \tag{5}$$

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

5 Empirical strategy

5.1 Response to corrections

To examine whether the data is consistent with Proposition 1, 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 farther 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 $\mu_i \equiv \sum_{k=1}^{N} \mu^k \mathbb{1}[i = k]$ are the 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.[23]

More specifically, given the random pairing of participants, the paired participant's gender is exogenous to the participant's unobservables. However, correction is not exogenous for two reasons: (i) correction can be correlated with the paired participant's ability, and the paired participant's ability can affect the participant's willingness to collaborate; (ii) the participant's personality – for example, overconfidence – affects their puzzle behavior, which in turn affects the paired participant's

23. 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.

behavior. To address the former point, I assume that $Contribution_j$ fully captures j's ability as perceived by i through j's puzzle moves (not true ability). This assumption is reasonable if we think participants' willingness to collaborate increases with the partners' contributions to the puzzle, which is consistent with the fact that participants can partially observe their partners' ability. To address the latter point, I add individual fixed effects: because j's unobservables are exogenous to i's unobservables and all i can observe about j is j's gender and puzzle behavior (correction and contribution), conditional on these observables about j, whether i selects j as their collaborator is an outcome of a particular paring which is random.

Also, as discussed in the theoretical framework (Section 4), good and bad corrections only have a signaling effect on j's ability after controlling for contributions; if i can fully observe j's ability, good and bad corrections convey no information that a rational agent cares about.

5.2 Heterogeneity by participants' ability

To examine whether the data is consistent with Proposition 2, 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 an above-median number of puzzles in part 1 in a session they participated in, 0 otherwise.

Other variables are as defined in equation 6.

5.3 Heterogeneity by partners' gender

To examine whether the data is consistent with Proposition 3, 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 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)
Good correction	-0.208***	-0.238***		-0.204***		-0.229***		-0.168***
	(0.028)	(0.030)		(0.024)		(0.033)		(0.036)
Bad correction	-0.518***	-0.508***		-0.100***		-0.172***		-0.011
	(0.031)	(0.034)		(0.036)		(0.047)		(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	0.002	0.004	0.016	0.016
	(0.016)	(0.017)	(0.014)	(0.014)	(0.018)	(0.018)	(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		0.347		0.016
Baseline mean	0.780	0.780	0.780	0.780	0.783	0.783	0.778	0.778
Baseline SD	0.414	0.414	0.414	0.414	0.413	0.413	0.416	0.416
Adj. R-squared	0.104	0.100	0.334	0.335	0.365	0.369	0.306	0.306
Observations	3180	3180	3180	3180	1670	1670	1510	1510
Individuals	464	464	464	464	244	244	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 the partner's contribution and individual fixed effects, and column 2 excludes the partner's contribution. Column 3 combines good and bad corrections as a single dummy variable. Columns 5-6 present the corresponding results for women and columns 7-8 for men. The p-values (F-test) for the differences of the coefficient across columns: 0.333 for any correction in column 5 and column 7, 0.178 for good correction in column 6 and column 8, and 0.184 for bad correction in column 6 and column 8. Baseline mean and standard deviation are participants' 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%.

6 Results

6.1 Response to corrections

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 while column 2 excludes the partner's contribution. Column 3 combines good and bad corrections 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 the 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 percentage points 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 0.271 more negative than the coefficient estimates on good corrections (p-value < 0.01).

Corroborating this, as column 3 shows, the coefficient estimate on the partner's contribution,

0.083, is positive and both quantitatively and statistically highly significant (p-value < 0.01). This suggests that participants are 8.3 percentage points more willing to collaborate with partners who make one more good move. This is true both for women (column 5, 0.090 with p-value < 0.01) and men (column 7, 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 are more willing to collaborate with partners who contributed more.

The coefficient estimate on any correction in column 3, -0.198, is negative and both quantitatively and statistically highly significant (p-value < 0.01). This suggests that participants are 19.8 percentage points less willing to collaborate with those who made one or more corrections. To offset this effect, a partner's contribution has to increase by 0.79 standard deviations.[24] The corresponding coefficient estimates for women are -0.237 (column 5, p-value < 0.01) and -0.152 for men (column 7, p-value < 0.01). Thus, participants are less willing to collaborate with a person who corrected their move.

This negative response to a correction 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 estimates for women are -0.229 (column 6, p-value < 0.01) and -0.168 for men (column 8, p-value $<$ 0.01).

The coefficient estimate on bad correction in column 4, -0.100, is also negative and both quantitatively and statistically significant (p-value < 0.01). However, the magnitude is smaller than the coefficient estimate on good correction, with a difference of -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 only -0.011 (column 8, 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, responses to corrections depend on participant's belief about their own ability relative to their partners. People are in general overconfident, albeit that men are more overconfident (Croson and Gneezy 2009). Thus, these behaviors may not be irrational.

6.2 Heterogeneity by participants' ability

Table 4 shows that the negative response to corrections we observed in the previous subsection is likely to be irrational: the table presents the regression results of equation 8. As in Table 3, columns 1-2 include all participants' willingness to collaborate. Columns 3-4 show the corresponding results for women and columns 5-6 for men.

24. 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 and 2.87 for partners faced by men: (2.73+2.87)/2=2.80.

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

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. The p-values (F-test) for the differences of the coefficient across columns: 0.810 for any correction in column 3 and column 5, 0.944 for good correction in column 4 and column 6, 0.137 for bad correction in column 4 and column 6, 0.073 for any correction times high ability in column 3 and column 5, 0.057 for good correction times high ability in column 4 and column 6, and 0.409 for bad correction times high ability in column 4 and column 6. Baseline mean and standard deviation are participants' 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%.

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 it is more negative for men (column 5, p-value < 0.05). Thus, high-ability participants, in particular men, dislike receiving corrections more than low-ability participants.

It is not efficiency deteriorating or irrational 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, but it is more negative and statistically significant (p-value < 0.05) for men (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 participants respond negatively to good corrections, with men responding more negatively. This suggests that a negative reaction to corrections is likely to be irrational: as discussed at the beginning of this section, high-ability participants should be able to distinguish between good and bad corrections and should respond less negatively to good corrections and more negatively to bad corrections than low-ability participants, as the rational agent benchmark in Proposition 2 suggests. However, what we see here is the opposite.

Aside from these main results, comparing the coefficient estimates on any correction in columns 3 and 5, high ability male participants dislike receiving corrections more than high ability female participants, although statistically significant only at 10% level. This effect mainly comes from good correction, as we can see from comparing the coefficient estimates on any correction in columns 4 and 6. This gender difference could be due to that men are more overconfident than women.

6.3 Heterogeneity by partners' gender

Table 5 presents the regression results of equation 8. As in 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 the 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 results suggest that participants – both women and men – do not underestimate women's contributions when selecting a collaborator. In other words, they correctly believe that women and men are equally good at solving the puzzle.

In column 1, the coefficient estimate 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 an asymmetric response: the coefficient estimate on the interaction between good correction and female partner is negative although statistically insignificant, but the coefficient estimate 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: as shown in 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

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. The p-values (F-test) for the differences of the coefficient across columns: 0.924 for any correction in column 3 and column 5, 0.732 for good correction in column 4 and column 6, 0.947 for bad correction in column 4 and column 6, 0.253 for any correction times female partner in column 3 and column 5, 0.061 for good correction times female partner in column 4 and column 6, and 0.274 for bad correction times female partner in column 4 and column 6. Baseline mean and standard deviation are participants' 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%.

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 – or be led by women – but they are okay with that women make mistakes. As referred to in the introduction, several studies document men's aversion to being led by women (Abel 2022; Chakraborty and Serra 2022; Husain, Matsa, and Miller 2021).

7 Robustness checks

7.1 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 the pair is unlikely to solve the puzzle and good and bad corrections may not be meaningful.

7.2 Excluding rounds 6 and 7

Remember that in rounds 6 and 7, participants' willingness to collaborate is lower, they correct others more, and they are less likely to solve the puzzle, as shown in Figure 5 in section 3. As discussed in section 3, these are all outcomes of a particular pair independent of the type of the partner, but one may wonder whether these rounds are driving the results.

7.3 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 minds in determining whether to collaborate with their partners.

7.4 Robustness results

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).

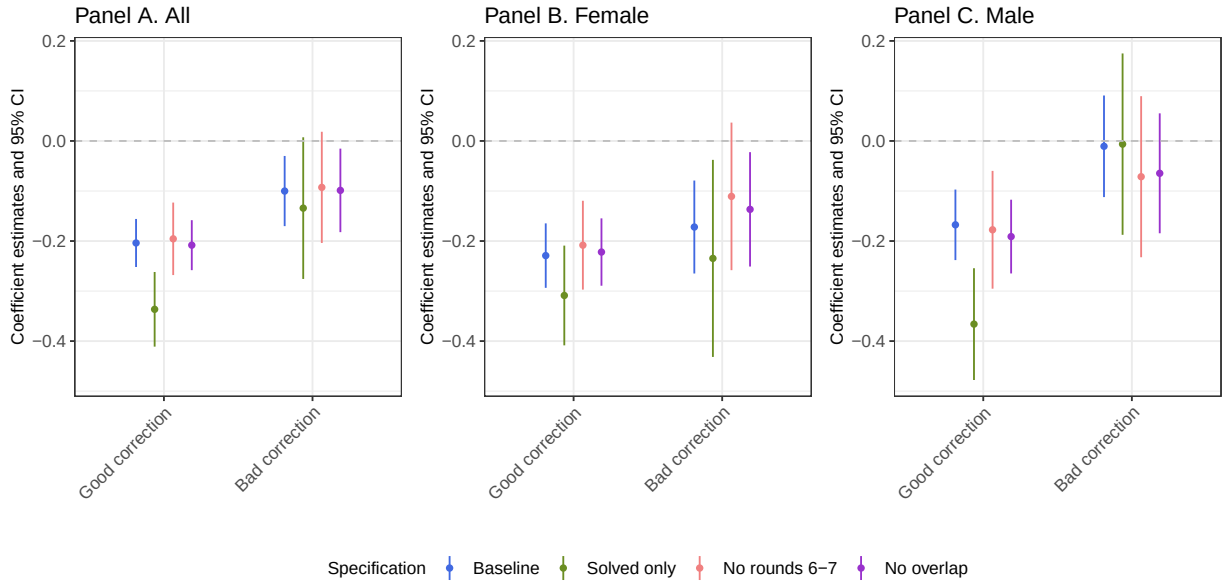
The main coefficients of interest for equation 6 are good and bad corrections. Looking at Figure 6, we see that most coefficient estimates are close to the main estimates. The estimates are more negative for good correction when the sample is limited to solved puzzles only, but they are more in line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

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

The main coefficients of interest for equation 8 are the interactions between good correction and female partner and between bad correction and female partner. Looking at Figure 8, we again see most of the coefficient estimates are close to the main estimates. 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 the evidence that men react more negatively to women's good corrections.

Figure 6: Response to corrections: Robustness



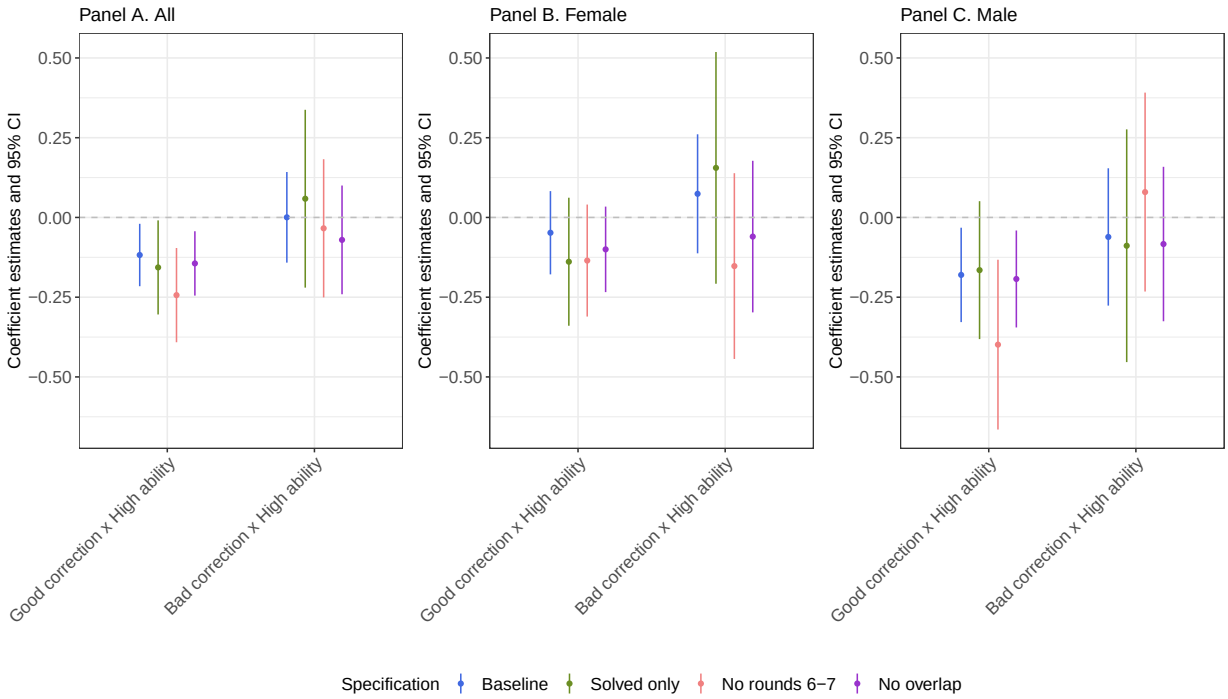
Notes: This figure plots the coefficient estimates and 95% confidence intervals of columns 4, 7, and 10 of Table 3 with solved puzzles only (the green dots and lines), with rounds 1-5 only (the red dots and lines), and with puzzles where only good or bad corrections occurred (the purple dots and lines).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

8 Conclusion

This paper demonstrates that people, including those with high ability, are less willing to collaborate with someone who has corrected them, even if the correction improved group performance. I also provide suggestive evidence that men respond more negatively to women's corrections that improve group performance but not to women's corrections that deteriorate group performance, presumably because men do not like women to correct their mistakes. On the contrary, women respond roughly the same to women's and men's corrections. 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.

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-solving ability is not informative of an ability relevant for the participants' work or study – it is not something they have been devoting much of their time to,

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



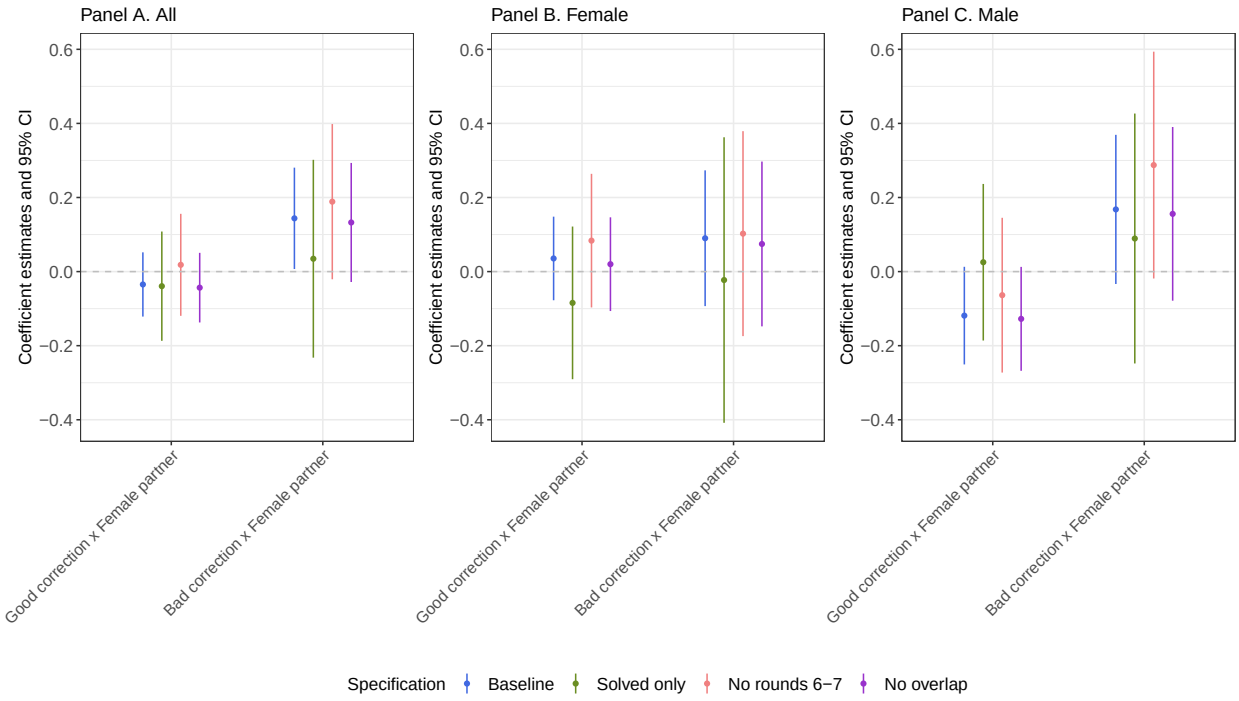
Notes: This figure plots the coefficient estimates and 95% confidence intervals of columns 2, 4, and 6 of Table 4 with solved puzzles only (the green dots and lines), with rounds 1-5 only (the red dots and lines), and with puzzles where only good or bad corrections occurred (the purple dots and lines).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

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 costs, using tasks that are more related to one's real-world ability, and having variation in seniority would be interesting extensions of this paper.

However, my experiment has 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, though 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's or master's students who are supposed to have a weaker gender bias than the general working population, due to their age and that they are presumably more aware of that gender bias is a bad thing. The first point relates to the takeaway of my results: it would be worth investigating whether a good workplace climate mitigates negative reactions to corrections. The second point relates to the study's external validity: women's corrections may receive stronger and more robust negative reactions in real workplace environments where people are older and possibly less educated.

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

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 the samples in these ways.

results are consistent with self-image concerns and information avoidance (Golman, Hagmann, and Loewenstein 2017).[25] For example, Kszegi (2006) finds that people avoid a difficult task when it reveals their ability. Corroborating this, Castagnetti and Schmacker (2022) find people select information that is 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. Regarding gender, evidence suggests that people respond differently to women's and men's feedback (Sinclair and Kunda 2000). A possible interpretation of my results is that receiving good corrections is a negative feedback, and accepting them damages people's self-image.[26]

25. Abelson (1986) is probably the first to propose this idea, who argues that people's "beliefs are like possessions" (p. 223).

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

References

Abel, Martin. 2022. "Do Workers Discriminate against Female Bosses?" *Journal of Human Resources.*

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. 2022. "Improving Workplace Climate in Large Corporations: A Clustered Randomized Intervention." *The Quarterly Journal of Economics.*

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

Blau, Francine D., and Lawrence M. Kahn. 2017. "The Gender Wage Gap: Extent, Trends, and Explanations." *Journal of Economic Literature* 55 (3): 789–865.

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

Born, Andreas, Eva Ranehill, and Anna Sandberg. 2022. "Gender and Willingness to Lead: Does the Gender Composition of Teams Matter?" *The Review of Economics and Statistics* 104 (2): 259–275.

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. 2022. "Protecting the Ego: Motivated Information Selection and Updating." *European Economic Review* 142:104007.

Chakraborty, Priyanka, and Danila Serra. 2022. *Gender and Leadership in Organizations: The Threat of Backlash.* Working Paper.

Chance, Zoë, and Michael I. Norton. 2015. "The What and Why of Self-Deception." *Current Opinion in Psychology,* Morality and Ethics, 6:104–107.

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.

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

Danz, David, Neeraja Gupta, Marissa Lepper, Lise Vesterlund, and K. Pun Winichakul. 2021. *Going Virtual: A Step-by-Step Guide to Taking the in-Person Experimental Lab Online.* Working Paper.

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.

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.

Folke, Olle, and Johanna Rickne. 2022. "Sexual Harassment and Gender Inequality in the Labor Market." *The Quarterly Journal of Economics* 137 (4): 2163–2212.

Gino, Francesca, Michael I. Norton, and Roberto A. Weber. 2016. "Motivated Bayesians: Feeling Moral While Acting Egoistically." *Journal of Economic Perspectives* 30 (3): 189–212.

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* 117 (1): 60–76.

Guo, Joyce, and María P. Recalde. 2022. "Overriding in Teams: The Role of Beliefs, Social Image, and Gender." *Management Science.*

Haeckl, Simone, and Mari Rege. 2022. *Effects of Supportive Leadership Behaviors on Employee Satisfaction, Engagement, and Performance: An Experimental Field Investigation.* Working Paper.

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

Husain, Aliza N., David A. Matsa, and Amalia R. Miller. 2021. *Do Male Workers Prefer Male Leaders? An Analysis of Principals Effects on Teacher Retention.* Working Paper.

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.

Kunda, Ziva. 1990. "The Case for Motivated Reasoning." *Psychological Bulletin* 108 (3): 480–498.

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.

Li, Jiawei, Stephen Leider, Damian Beil, and Izak Duenyas. 2021. "Running Online Experiments Using Web-Conferencing Software." *Journal of the Economic Science Association* 7 (2): 167–183.

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.

Zhao, Shuchen, Kristian López Vargas, Daniel Friedman, and Marco Antonio Gutierrez Chávez. 2020. *UCSC LEEPS Lab Protocol for Online Economics Experiments.* Working Paper.

Online Appendix

A Deviations from the pre-analysis plan

In the pre-analysis plan, I specified the following three hypotheses (rephrasing some terms to match the text; see the original phrasing in Online Appendix D):

H1 Men are less likely to collaborate with a woman than with a man who corrects them.

H2 The behavior conjectured in H1 leads to a suboptiomal collaborator choice.

H3 A mechanism that underlies the behavior conjectured in H1 is gender bias.

I presented the results for H1 and H2 in Table 5, but did not present the results for H3 because I could not detect a meaningful variation in the gender bias measure in my sample obtained from the six hostile and benevolent sexism questions used in Stoddard, Karpowitz, and Preece (2020), which they selected from Glick and Fiske (1996). These results are presented in Subsection A.1.

In contrast, I presented in the main text two non-pre-registered results because I found the patterns worth investigating and interesting. The first was people's unwillingness to collaborate with a person who corrects them, regardless of the corrector's gender, presented in Table 3. The second was heterogeneity of people's unwillingness by their puzzle ability to examine whether the results in Table 3 was due to their misunderstanding of good corrections as bad corrections, presented in Table 4.

In addition, I modified the definition of contribution from the one in the pre-analysis plan because there was truncation in the pre-registered contribution measure in more than 10% of the puzzle, and I thought the modified measure, which did not truncate, was more correct. Nonetheless, the same results hold when I use the original contribution measure, reported in Tables A2, A3, A4, and A5. 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.

A.1 Pre-specified analysis not reported in the main text

I estimate the following model with OLS.

$$\begin{aligned} Select_{ij} =& \beta_1 CorrectedGood_{ij} + \beta_2 CorrectedBad_{ij} + \beta_3 Female_j \\ &+ \beta_4 CorrectedGood_{ij} \times HighBias_i + \beta_5 CorrectedBad_{ij} \times HighBias_i \\ &+ \delta_1 Contribution_j + \delta_2 Contribution_j \times HighBias_i + \mu_i + \epsilon_{ij} \end{aligned} \tag{A1}$$

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 equation 6.

Table A1 presents the regression results of equation A1. 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 columns 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 is 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.

A.2 Results with the pre-registered contribution measure

Table A1: 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 A1. 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 participants' 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 A2: Response to corrections: The pre-registered contribution measure

Dependent variable:	Willing to collaborate (yes=1, no=0)							
Sample:	All				Female		Male	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Good correction	-0.208***	-0.238***		-0.272***		-0.304***		-0.230***
	(0.028)	(0.030)		(0.026)		(0.035)		(0.038)
Bad correction	-0.518***	-0.508***		-0.160***		-0.234***		-0.065
	(0.031)	(0.034)		(0.037)		(0.048)		(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	0.001	0.004	0.012	0.012
	(0.016)	(0.017)	(0.014)	(0.014)	(0.019)	(0.019)	(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		0.252		0.014
Baseline mean	0.780	0.780	0.780	0.780	0.783	0.783	0.778	0.778
Baseline SD	0.414	0.414	0.414	0.414	0.413	0.413	0.416	0.416
Adj. R-squared	0.104	0.100	0.309	0.314	0.320	0.330	0.300	0.300
Observations	3180	3180	3180	3180	1670	1670	1510	1510
Individuals	464	464	464	464	244	244	220	220

Notes: This table reports the same estimation results as Table 3 but with the pre-registered contribution measure specified in the pre-analysis plan, and shows that the results are robust to using the pre-registered measure. The p-values (F-test) for the differences of the coefficient across columns: 0.127 for any correction in column 5 and column 7, 0.064 for good correction in column 6 and column 8, and 0.385 for bad correction in column 6 and column 8. Baseline mean and standard deviation are participants' 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 A3: Response to corrections of high vs. low ability people: The pre-registered 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 pre-registered contribution measure specified in the pre-analysis plan, and shows that the results are robust to using the pre-registered measure. The p-values (F-test) for the differences of the coefficient across columns: 0.612 for any correction in column 3 and column 5, 0.526 for good correction in column 4 and column 6, 0.238 for bad correction in column 4 and column 6, 0.167 for any correction times high ability in column 3 and column 5, 0.069 for good correction times high ability in column 4 and column 6, and 0.296 for bad correction times high ability in column 4 and column 6. Baseline mean and standard deviation are participants' 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 A4: Response to corrections made by women vs. men: The pre-registered 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 pre-registered contribution measure specified in the pre-analysis plan, and shows that the results are robust to using the pre-registered measure. The p-values (F-test) for the differences of the coefficient across columns: 0.740 for any correction in column 3 and column 5, 0.858 for good correction in column 4 and column 6, 0.748 for bad correction in column 4 and column 6, 0.321 for any correction times female partner in column 3 and column 5, 0.117 for good correction times female partner in column 4 and column 6, and 0.218 for bad correction times female partner in column 4 and column 6. Baseline mean and standard deviation are participants' 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 A5: Response to corrections made by women vs. men: Heterogeneity by gender bias, the pre-registered 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 A1 but with the pre-registered contribution measure specified in the pre-analysis plan, and shows that the results are robust to using the pre-registered measure. Baseline mean and standard deviation are participants' 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%.

B Pros and cons of the quasi-laboratory format

The quasi-laboratory experimental format I use in this paper has both pros and cons over the standard online and physical laboratory format. Since it is relatively new, I discuss them below.

Over the 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 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 participants 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 hundred, but it weakens the connection between the experimenter and the participants and can induce attrition.

Over the 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 systems, 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.

The key drawback is that participants can search internet for real effort tasks and questions. Thus, the tasks and questions have to be non-internet searchable.

There are already a few studies that use a quasi-laboratory format, for example, Goeschl, Oestreich, and Soldà (2021). Also, there are several guidelines about how to conduct quasi-laboratory experiments (Danz et al. 2021; Li et al. 2021; Zhao et al. 2020).

C 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.

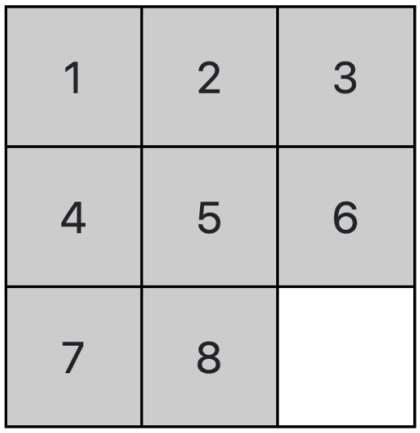
App: pt1

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

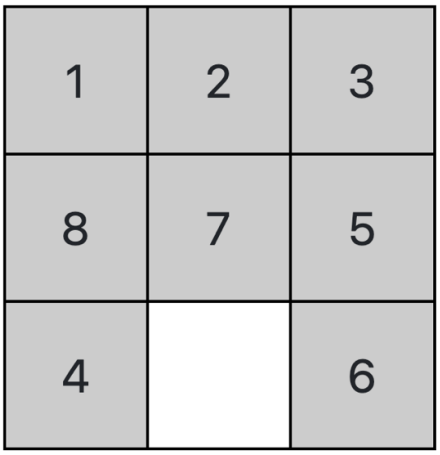
First, there are 2 moves you can make: move 2 right and move 3 up. Which moves should you make?

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.

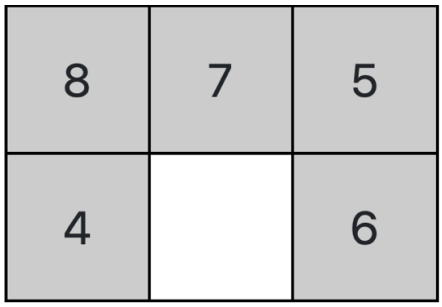
Example 3:

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

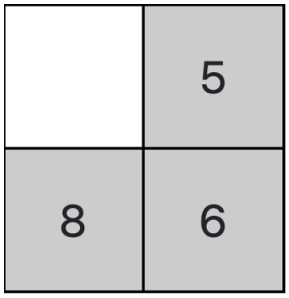
Now, if you ignore the top row which is already in the correct order, the only tile you can move is 7. So, let's move it to the left.

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

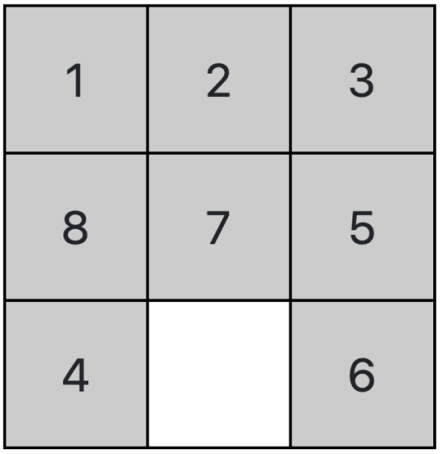
Now let's also keep the left column as is.



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.

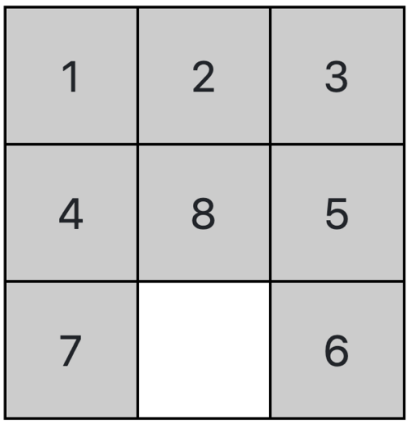


puzzle_3_0.png

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.

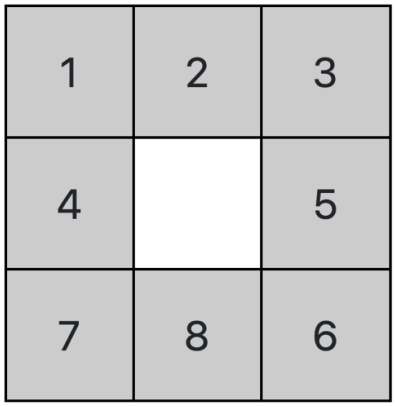
puzzle_3_3_good.png

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.

D Pre-analysis plan

Gender differences in the cost of contradiction
Pre-analysis plan

Yuki Takahashi

November 22, 2020

This document pre-specifies the main hypotheses, the experimental design, and the empirical specifications for a laboratory experiment that examines gender differences in the cost of contradiction. At the time this document is written, I ran 1 pilot session (with 16 participants) to make sure that the experimental design and procedure worked without any problems.

1 Main Hypotheses

H1: Men are less likely to work with a woman than with a man who contradicts them.
H2: The behavior conjectured in H1 leads to a suboptimal partner choice.
H3: A mechanism that underlies the behavior conjectured in H1 is gender bias.

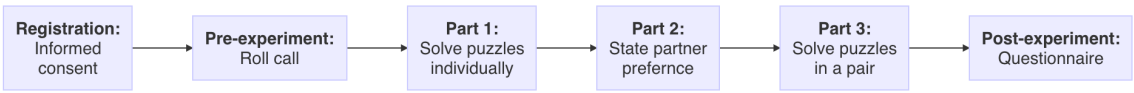
2 Design and procedure

The experiment will be computerized and conducted online with the University of Bologna's students in Italian. However, unlike standard online experiments, I will conduct the experiment as a "quasi-laboratory" where participants will be connected with the experimenter via Zoom throughout the experiment and listen to the instructions the experimenter will read out, ask questions to the experimenter via private chat, etc., just like the standard laboratory experiment. Their camera and microphone will be turned off throughout the experiment except when the experimenter calls their name at the beginning of the experiment (explained later).

Based on the power simulation in appendix A, I will recruit approximately 450 participants (225 female and 225 male). Each session will consist of a multiple of 8 participants and is expected to last for 1 hour. The average total payment per participant will be 10€, the maximum 25€, and the minimum 2€, all including the 2€ participation fee.

I use Isaksson (2018)'s 3x3 sliding puzzle as the real effort task for this experiment and define the difficulty (the number of moves away from the solution), good moves (a move that reduces the number of moves away from the solution), and bad moves (a move that increases the number of moves away from the solution) by the Breadth-First Search algorithm.

Figure 1: Flowchart of the experiment



The experiment will consist of 3 parts as summarized in figure 1. The details are below:

Registration

1. Upon receiving the invitation email to the experiment, participants will register for a session they want to participate in and upload their ID documents as well as a signed consent form. I will recruit a few more participants than I will need for a given session in case some participants would not show up to the session.

Pre-experiment

2. On the day and the time of the session they have registered, the participants will enter the Zoom waiting room. They receive a link to the oTree virtual room and enter their first name, last name, and their email they have used in the registration. They also draw a virtual coin that is numbered from 1 to 40.
3. Then I admit participants to the Zoom meeting room one by one and rename them by the first name they have entered on the oTree. If there is more than one participant with the same first name, I will add a number after their first name (e.g. Giovanni2).
4. After admitting all the participants, I will do roll call: I will call participants' first names and ask them to respond via microphone to ensure other participants that the called participants' first names correspond to their gender. If there are more participants than I would need to run the session, I will draw random numbers from 1 to 40 and ask those who drew the coins with the same number to leave. Those who will leave the session will receive the participation fee.

Part 1: Individual round

5. Participants will work on the puzzle individually with an incentive (0.2€ per puzzle solved). They can solve as many puzzles as possible with increasing difficulty (but maximum of 15 puzzles) in 4 minutes. This part will familiarize them with and measure their ability to solve the puzzle. The ability is measured by the number of puzzles they solve.

Part 2: Partner preference elicitation

6. Participants will be told the rules of part 3 and state their partner preference. This part will proceed as follows: participants will be grouped into 8 participants based on their ability similarity, then each participant will be randomly matched with another participant in the same group and solve 1 puzzle together by alternating their move. Which participant will make the first move will be randomized and this will be told to both participants. If they cannot solve the puzzle within 2 minutes, they will finish the puzzle without solving it. Reversing the matched participant's move will be used as the measure of contradiction. The matched participant's first name will be displayed on the computer screen throughout the puzzle to subtly inform that participant's gender. Each participant's contribution to a given puzzle is measured as defined in appendix C.
7. Once they finish the puzzle, participants will state whether they want to work with the matched participant (yes/no), which will be used as the measure of their partner preference.

Then they will be randomly re-matched with another participant with a perfect stranger algorithm and repeat point 6 with a different puzzle with the same difficulty and state their partner preference.

8. After all the participants solve the puzzle with all the other participants in the same group and state their partner preference, participants are matched according to the following algorithm:
   (a) 1 participant is randomly chosen
   (b) if they have a match (both them and the other person state "yes" when they are matched) they will work together in part 3
   (c) if they have more than 1 matches, 1 of the matches is randomly chosen
   (d) the match is excluded and (a)-(c) is repeated until there is no match
   (e) if some participants are still left unmatched, they are matched randomly

Part 3: Group round

9. The matched participants will work together on the puzzles by alternating their move for 12 minutes and earn 1€ for each puzzle solved. Which participant will make the first move will be randomized at each puzzle and this will be told to both participants as in part 2. They can solve as many puzzles as possible with increasing difficulty (but maximum of 20 puzzles).

Post-experiment

10. Participants will answer a short questionnaire which consists of (i) the 6 hostile and benevolent sexism questions in Stoddard, Karpowitz, and Preece (2020) which is originally from Glick and Fiske (1996) and measure gender bias,[1] and (ii) their basic demographic information and what they have thought about the experiment (see appendix B for the questions asked). I will ask them these questions in this order.
11. After participants answer all the questions, I will tell them their earnings and let them leave the virtual room and Zoom. They will receive their earnings via PayPal.

3 Specification

Test of H1 I test H1 by estimating the following OLS regression using male participants' partner preference observations elicited in part 2. I call participants who state their partner preference as decision-makers, participants who are evaluated by the decision-makers as participants:

$$\begin{aligned} Prefer_{ij} = \beta_1 Contradict_{ij} * Female_j + \beta_2 Contradict_{ij} + \beta_3 Female_j \\ + \delta Contribute_{ij} + IndividualFE_i + \epsilon_{ij} \end{aligned} \tag{1}$$

- $Prefer_{ij} \in \{0, 1\}$: a dummy variable indicating whether decision maker i preferred participant j as their partner.
- $Contradict_{ij} \in \{0, 1, ...\}$: the number of times j reverses i's move.

1. The Italian translation is from Manganelli Rattazzi, Volpato, and Canova (2008) and Rollero, Glick, and Tartaglia (2014). I score the participants' answer following Stoddard, Karpowitz, and Preece (2020) (assign 0 to strongly disagree and 4 to strongly agree, take the arithmetic average of all the 6 questions, and divide it by 24).

- $Female_j \in \{0, 1\}$: an indicator variable equals 1 if participant j is female, 0 otherwise.
- $IndividualFE_i$: fixed effects for decision-maker i. This is necessary for identification for 2 reasons. First, i's unobserved characteristics can affect both j's puzzle play (j's contradiction and contribution) and the probability that i prefers j as a partner. Second, the wealth effect is different across i because each i can earn a different amount in part 1.
- $Contribute_{ij} \in [0, 1]$: participant j's contribution to a puzzle played with decision-maker i as defined in appendix C. This is necessary for identification so that I can compare women and men who contradict i and make the same contribution. I add this variable as a linear term because the outcome must be increasing in j's contribution.

β_1 compares decision-makers' partner preference for female vs male participants who make the same number of contradictions and tests H1:

- $\beta_1 < 0$: men are less likely to work with a woman than with a man who contradicts them (so yes to H1).
- $\beta_1 > 0$: men are more likely to work with a woman than with a man who contradicts them (so no to H1).
- $\beta_1 = 0$: men are neither more nor less likely to work with a woman than with a man who contradicts them (so no to H1).

Test of H2 To test H2, I separate the effect of good contradictions in equation 1 by estimating the following OLS regression using the same sample as test of H1.

$$\begin{aligned} Prefer_{ij} =& \beta_1 Contradict_{ij} * Female_j + \beta_2 Contradict_{ij} + \beta_3 Female_j \\ &+ \beta_4 ContradictGood_{ij} * Female_j + \beta_5 ContradictGood_{ij} \\ &+ \delta Contribute_{ij} + IndividualFE_i + \epsilon_{ij} \end{aligned} \tag{2}$$

- $ContradictGood_{ij} \in \{0, 1, ...\}$: the number of times j reverses i's bad move.

other variables are as defined in equation 1.

β_4 picks up the part of β_1 in equation 1 that comes from j's good contradiction and tests H2:

- $\beta_4 < 0$: the behavior conjectured in H1 leads to a suboptimal partner choice (so yes to H2).
- $\beta_4 > 0$: the behavior conjectured in H1 leads to an optimal partner choice (so no to H2).
- $\beta_4 = 0$: the behavior conjectured in H1 leads to neither a suboptimal nor an optimal partner choice (so no to H2).

Test of H3 To test H3, I interact the contradictions, participants' gender, and their interaction with decision-makers' gender bias in 1 by estimating the following OLS regression using the same sample as test of H1.

$$\begin{aligned} Prefer_{ij} =& \beta_1 Contradict_{ij} * Female_j + \beta_2 Contradict_{ij} + \beta_3 Female_j \\ &+ \beta_4 Contradict_{ij} * Female_j * StrongerBias_i + \beta_5 Contradict_{ij} * StrongerBias_i \\ &+ \beta_6 Female_j * StrongerBias_i + \delta Contribute_{ij} + IndividualFE_i + \epsilon_{ij} \end{aligned} \tag{3}$$

- $StrongerBias_i \in \{0, 1\}$: an indicator variable equals 1 if decision-maker i's gender bias measured by the 6 hostile and benevolent sexism questions in the post-experimental questionnaire is above median of all the male decision-makers, 0 otherwise.

other variables are as defined in equation 1.

β_4 tests whether the behavior conjectured in H1 is stronger among decision-makers with stronger gender bias and tests H3:

- $\beta_4 < 0$: the behavior conjectured in H1 is stronger among decision-makers with stronger gender bias (so yes to H3).
- $\beta_4 > 0$: the behavior conjectured in H1 is weaker among decision-makers with stronger gender bias (so no to H3).
- $\beta_4 = 0$: the behavior conjectured in H1 is neither stronger nor weaker among decision-makers with stronger gender bias (so no to H3).

Standard error adjustment Because the treatment unit is i, I cluster standard error at i. Although the same individual appears twice (once as i and once as j), j is passive in preference elicitation.

Unsolved puzzles I include pairs who could not solve the puzzle.

Notes about the tests of H2 and H3 Interpreting the tests for H2 and H3 may require cautions. First, both tests are likely to be underpowered because they further split the effect of H1 for which the sample size is determined. Second, only for the test of H3, participants may not answer the gender bias questions honestly because gender is a socially sensitive issue, so the test may not be able to detect the effect even if H3 is true.

References

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

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

Manganelli Rattazzi, Anna Maria, Chiara Volpato, and Luigina Canova. 2008. "L'Atteggiamento ambivalente verso donne e uomini: Un contributo alla validazione delle scale ASI e AMI. [Ambivalent attitudes toward women and men: Contribution to the validation of ASI and AMI scales.]" *Giornale Italiano di Psicologia* 35 (1): 217–243.

Rollero, Chiara, Peter Glick, and Stefano Tartaglia. 2014. "Psychometric properties of short versions of the Ambivalent Sexism Inventory and Ambivalence Toward Men Inventory." *TPM-Testing, Psychometrics, Methodology in Applied Psychology* 21 (2): 149–159.

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

Appendix A Power simulation

I estimate the number of participants I have to recruit to achieve 80% power for the test of H1 via Monte Carlo simulation.

I assume the following data generating process:

$$\begin{aligned} Prefer^*_{ij} = b_0 + b_1 Contradict_{ij} * Female_{ij} + b_2 Contradict_{ij} + b_3 Female_{ij} \\ + \delta Contribute_{ij} + \sum_{k=1}^{3} \gamma^k \mathbb{1}(a_i = k) + \sum_{k=1}^{3} \theta^k \mathbb{1}(m_i = k) + e_{ij} \\ (i = 1, ..., N; j = 1, ..., 7) \end{aligned} \tag{A1}$$

where each variable is drawn from the following distribution:

- $Contradict_{ij} \sim Pois(0.1\frac{L}{2} + 0.02(m_i - 1)\frac{L}{2})$ (10% of moves were reversed following Isaksson (2018); the meaner the decision-maker, the more likely they receive a contradiction)
- $Female_{ij} \sim^{iid} Bernoulli(0.5)$ (a matched participant is female by 50% chance)
- $Contribute_{ij} \sim TN(0.5 - 0.1(a_i - 1.5), 0.05, 0, 1)$ (a matched participant's contribution which negatively depends on the decision-maker's ability)
- $a_i \sim^{iid} Unif\{1, 3\}$ (the decision-maker's ability)
- $m_i \sim^{iid} Unif\{1, 3\}$ (the decision-maker's meanness)
- $e_{ij} \sim^{iid} N(0, \sigma^2)$ (large sample approximation)
- $Prefer_{ij} = \mathbb{1}(Prefer^*_{ij} > 0)$

Each parameter is defined as follows:

- $b_0 = 0$ (so that the unconditional probability that the decision-maker chooses a matched participant is 50%)
- $b_1 = MDE$
- $b_2 = MDE$ (being contradicted by a female participant reduces the probability of choosing that participant as a partner twice as much as being contradicted by a male participants)
- $b_3 = 0$ (the decision-maker has no underlying gender bias)
- $\delta = 0.2$ (this is the main determinant of partner preference: the higher a matched participant's contribution, the higher the probability that the decision-maker chooses them as a partner)
- $\gamma^k = -0.02 * (k - 1.5)$, k=1,2,3 (the higher the decision-maker's ability, the lower the probability that the decision-maker chooses a matched participant as a partner)
- $\theta^k = -0.02 * (k - 1.5)$, k=1,2,3 (the meaner the decision-maker, the lower the probability that the decision-maker chooses a matched participant as a partner)
- $\sigma = 0.1$

where L is total number of moves the decision-maker and a matched participant take to solve a puzzle, which I assume to be 15 (7.5 moves by the decision-maker). However, I also set it to 10 (5 moves by the decision-maker) for robustness check. $MDE = -0.02$ is my baseline assumption (being contradicted once reduces the probability of choosing a matched participant by the same degree as when the matched participant's contribution is 0.1 lower), but I also set it to -0.01 for robustness check, -0.03 to see what happens in a more optimistic scenario, and 0 to check that

type I error rate is kept at 5% and that the estimated ATE is 0 when there is no underlying effect.

Thus, I estimate equation 1 with the sample drawn from equation A1 for $MDE \in \{0, -0.01, -0.02, -0.03\}$, $L \in \{15, 10\}$, and $N \in [50, 300]$. I draw 1000 independent sample.

Power is defined as the number of times the t-test rejects β_1 at 5% significance level (two-tailed) divided by the number of samples I draw:

$$Power(N, MDE, L) = \frac{\#Rejections(N, MDE, L)}{\#Draws} \tag{A2}$$

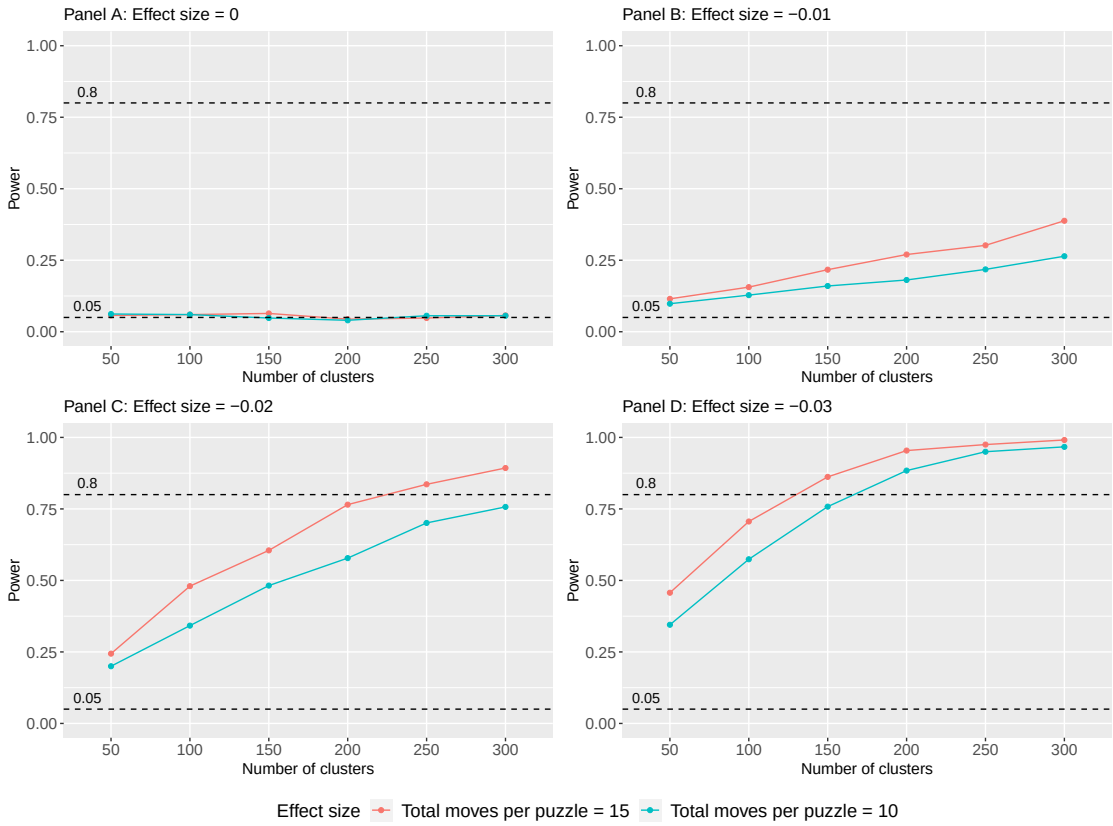
ATE is defined as the average of $\hat{\beta}_1$ across draws (its dependence on L is due to the non-linearity of the data generating process):

$$ATE(MDE, L) = \frac{\sum_{r=1}^{\#Draws} \hat{\beta}_1^r(MDE, L)}{\#Draws} \tag{A3}$$

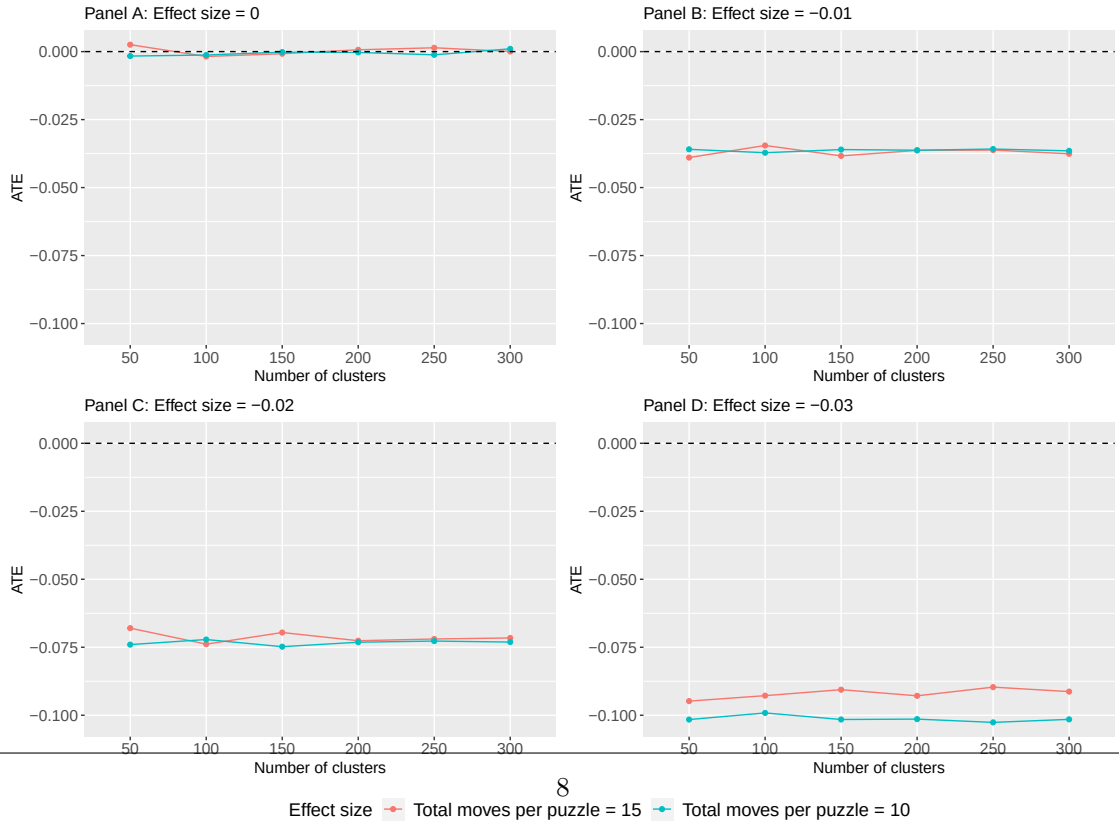
The results are presented in figure A1, which suggests that I need to recruit about 450 participants (so that I could have 225 clusters for testing H1). First, in the baseline scenario with $L = 15$, I can achieve about 80% power. Second, even under a tougher scenario where $L = 10$, I can still achieve about 60% power. The type I error rate is kept at 5%. ATE is larger than b_1 in magnitude because the data generating process is non-linear, but is 0 when the underlying effect size is 0. However, the power is very sensitive to the underlying effect size: if $MDE = -0.01$, I will likely not be able to detect the effect. If $MDE = -0.03$, on the other hand, my test is very high-powered: the power is close to 100% that I will almost always be able to detect the effect.

Figure A1: Estimated power and ATE (# draws=1000, $\alpha = 0.05$ two-tailed)

(a) Estimated power



(b) Estimated ATE



Appendix B Questions asked in the questionnaire

English version

- Your age: [Integer]
- Your gender: [Male, Female]
- Your region of origin: [Northwest, Northeast, Center, South, Islands, Abroad]
- Your major: [Humanities, Law, Social Sciences, Natural Sciences/Mathematics, Medicine, Engineering]
- Your 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]

Italian translation

- Età: [Integer]
- Sesso: [Uomo, Donna]
- Regione di origine: [Nord-Ovest, Nord-Est, Centro, Sud, Isole, Estero]
- Campo di studi principale: [Studi umanistici, Giurisprudenza, Scienze sociali, Scienze naturali/Matematica, Medicina, Ingegneria]
- Tipo di corso: [Laurea, Laurea Magistrale/Post-Laurea, Ciclo Unico (1 °, 2 ° o 3 ° anno), Ciclo Unico (4 ° anno o oltre), Dottorato]
- Cosa pensi di questo studio? [Textbox]
- C'era qualcosa di poco chiaro o di confuso in questo studio? [Textbox]
- I puzzle erano difficili? [Difficili, Abbastanza difficili, Giusto, Abbastanza facili, Facili]
- Hai qualche altro commento? (opzionale) [Textbox]

Appendix C Calculation of contribution

Following Isaksson (2018), I define a participant's contribution to a given puzzle in part 2 as follows:

$$\text{Player i's contribution} = \frac{P_i}{P_i + P_j} \in [0,1],\ i,j = 1,2,\ i \neq j \tag{C1}$$

$$P_i = \max\{\text{i's \# good moves} - \text{i's \# bad moves}, 0\}\ i = 1,2 \tag{C2}$$

If $P_i = 0$ and $P_j = 0$, I define both i's and j's contribution to 0.