Gender Differences in the Cost of Corrections in Group Work

Yuki Takahashi*

Click here for the latest version

August 23, 2022

Abstract

Collaboration is an integral component of workplace environments and part of collaboration also involves correcting one's colleagues. Using a quasi-laboratory experiment, I study whether people dislike collaborating with someone who corrects them and whether the dislike is stronger when that person is a woman. I find 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. In addition, I find suggestive evidence that men respond more negatively to women's corrections and that men's beliefs about gender differences in abilities cannot explain this differential response. These findings suggest that a behavioral bias distorts the optimal selection of talents and penalizes those who correct others' mistakes, and the distortion may be stronger when women correct men.

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

*Department of Economics, European University Institute. 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, 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 Bologna, and many other people. Tommaso Batistoni, Philipp Chapkovski, Christian König genannt Kersting, and oTree help & discussion group kindly answered my questions about oTree programming; in particular, my puzzle code was heavily based on Christian's code. Francesca Cassanelli, Natalia Montinari, and Ludovica Spinola helped me to write experimental instructions in Italian. Michela Boldrini and Boon Han Koh conducted the quasi-laboratory experiments ahead of me and kindly answered my questions about the implementations. Lorenzo Golinelli provided excellent technical and administrative assistance. This study was pre-registered with the OSF registry (https://osf.io/tgyc5) and approved by the IRB at the University of Bologna on November 3, 2020 (ref. no. 262643).

1 Introduction

Collaboration is the core element of the production process in the workplace, as most workplaces require group work (Jones 2021; Lazear and Shaw 2007; Wuchty, Jones, and Uzzi 2007). However, workplace interactions involve correcting one's colleagues: hiring committee members can have different opinions about the best candidates among the applicants, and co-authors in a research project can have conflicting views about the best experimental design or identification strategy. If people dislike being corrected, corrections can damage the collaborative relationship. This potential interpersonal friction can be detrimental to group efficiency because workplace climate is an important determinant of productivity (Alan, Corekcioglu, and Sutter 2021; Edmans 2011; Guiso, Sapienza, and Zingales 2015). In addition, women may experience stronger interpersonal friction because people, especially men, sometimes play down women to protect their self-image when women criticized them (Sinclair and Kunda 2000), which can contribute to the gender gap in labor market outcomes (Blau and Kahn 2017).

This paper studies whether people dislike collaborating with someone who corrects them and whether the dislike is stronger when that person is a woman. Answering this question using secondary data poses two challenges. First, group formation is not random, and group corrections are endogenous. Second, different corrections are not necessarily comparable to each other.[1]

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, I define corrections such that researchers can track the quality of corrections objectively. Specifically, participants are allocated to a group of eight and solve one collaborative task with each of the other group members, one after another (sequentially). Each time participants finish the task, they state whether they would like to collaborate with their current partner for the same task in the subsequent stage of the experiment, which is the main source of earnings. This gives a strong incentive for participants to select as good a collaborator as possible. The order of the group members with whom participants solve the task is randomized. As a collaborative task, I use Isaksson (2018)'s number-sliding puzzle, which allows me to calculate an objective measure of each participant's contribution to the collaborative task and to classify each move as good (move the puzzle closer to the solution) or bad (move the puzzle further away from the solution). I define a correction as reversing a group member's move; this gives us a measure that is comparable across different participants, and allows us to classify corrections as either good or bad.

I find that people understand the notion of good and bad moves; the higher one's contribution to solving the puzzle, the more likely it is that 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 contribution, people 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 people misunderstood good corrections as bad ones, because high ability people – who should be

1. I define collaboration as working with others toward the same goal, and correction as overriding what others said or did.

better able to identify good and bad corrections – also respond negatively to corrections. Thus, the negative response is likely to be irrational. Although only suggestive, I also find evidence that men respond more negatively to women's good corrections: men may dislike women correcting their mistakes, which is consistent with the recent studies that men dislike being led by women (Abel 2022; Chakraborty and Serra 2022; Husain, Matsa, and Miller 2021). This finding is unlikely to be due to men's beliefs about the differences in women's and men's abilities in the puzzle: women and men contribute equally well to the puzzle, and neither women nor men 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 men may exhibit stronger bias when women correct them.

This paper's contribution is twofold. First, it contributes to the literature on the workplace climate and productivity by showing that interpersonal frictions can distort group efficiency, and frictions may have a stronger effect on women. My findings complement Alan, Corekcioglu, and Sutter (2021), who find that a better workplace climate increases worker satisfaction and the degree of mutual reciprocation while reducing toxic competition and worker turnover and argue that improved manager-worker relationships are the likely mechanism. Aside from Alan et al., my findings also relate to the organizational economics literature. It finds that firms with high employee satisfaction exhibit higher stock prices (Edmans 2011) and that a firm performs better when its workers perceive their managers as trustworthy and ethical (Guiso, Sapienza, and Zingales 2015). Further, I show that the same environment can affect women 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.[2]

Second, this paper contributes to the literature on differential treatment of women's opinions by showing that women's corrections may 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.

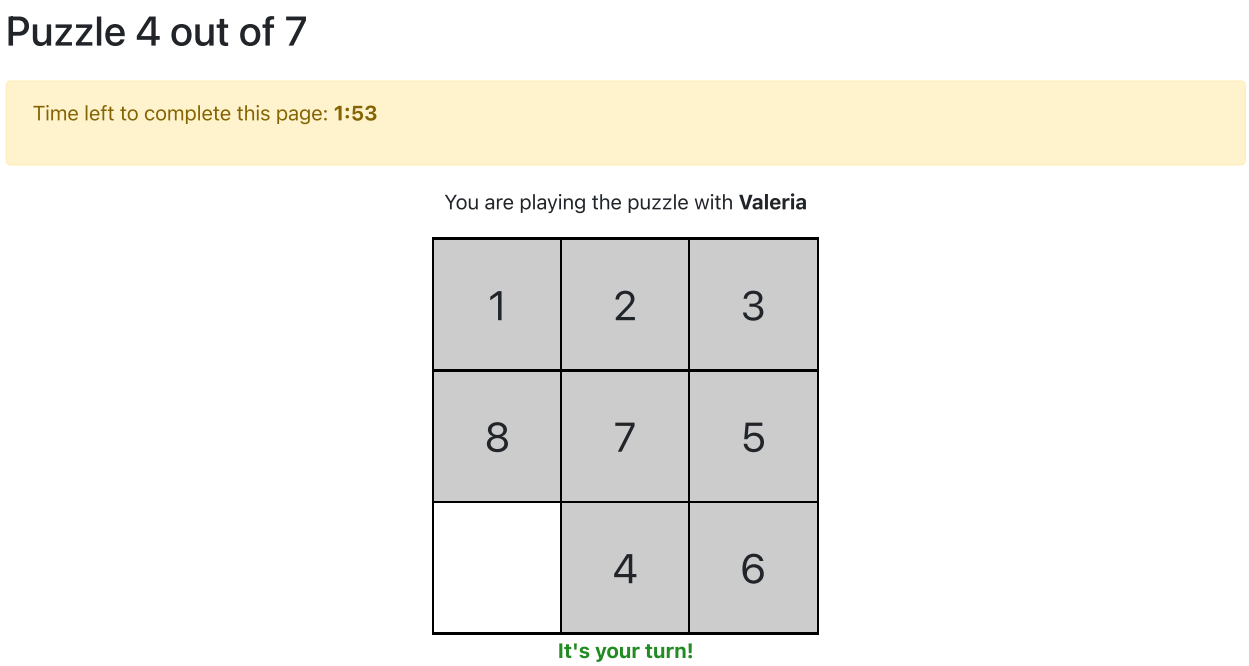
The remainder of the paper proceeds as follows. In section 2, I describe the experimental design, procedure, and implementation. Next, I describe the data obtained from the experiment in section 3. Then, I provide a simple theoretical framework to show how a rational agent would behave in section 4. Afterward, I proceed to empirical analysis: I present 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. Folke and Rickne (2022) also find the opposite: men in female-dominant jobs receive more harassment.

2 Experiment

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

Figure 1: Puzzle screen



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

Collaborative task As 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.[3] This puzzle has nice mathematical properties: I can define the puzzle difficulty and classify a given move as either good or bad by the Breadth-First Search algorithm.[4] From the number of good and bad moves one makes, I can calculate individual contributions to the task; I measure it by net good moves, the number of good moves minus the number of bad moves an individual makes in a given puzzle.

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

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

I can also determine the quality of corrections of different participants objectively and comparably.[5] 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).

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

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

2.1 Design and procedure

Registration

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

Pre-experiment

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

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

After admitting all the participants to the Zoom meeting room, I do roll call, a way to reveal participants' gender to other participants without making gender salient (Bordalo et al. 2019; Coffman, Flikkema, and Shurchkov 2021). Specifically, I take attendance by calling each participant's first name one by one and asking them to respond via microphone. This process ensures other participants that the called participant's first name corresponds to their gender. If there are more participants than I would need for the session (I need 16 participants), I draw random numbers

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

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

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

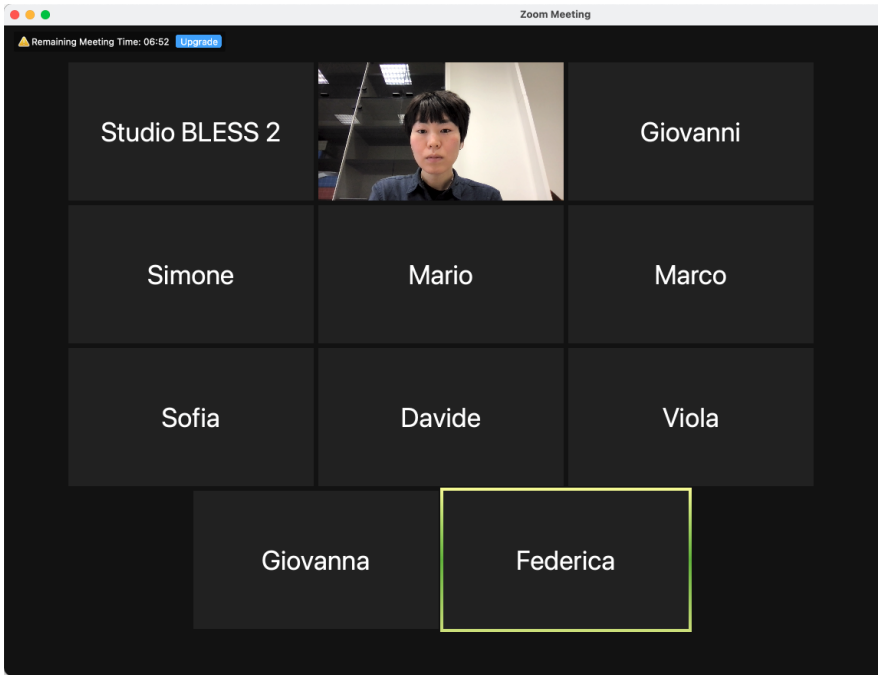
8. I recruit a few more participants than I would need for a given session in case some participants would not show up to the session.

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

from 1 to 40 and ask those who drew the coins with the same number to leave.[10] Those who leave the session receive the 2€ show-up fee. Figure 2 shows a Zoom screen participants would see during the roll call (the person whose camera is on is the experimenter; participants would see this screen throughout the experiment, but the experimenter's camera may be turned off).

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

Figure 2: Zoom screen



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

Part 1: Individual practice stage

Participants work on the puzzle individually with an incentive (0.2€ for each puzzle they solve). They can solve as many puzzles as possible with increasing difficulty (maximum 15 puzzles) in 4 minutes. After the 4 minutes, 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 given by the number of puzzles they solve.

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

Part 2: Collaborator selection stage

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

Second, participants are paired with another randomly chosen participant in the same group and solve one puzzle together by alternating their moves. The participant who makes the first move is drawn at random, and both participants know this first-mover selection criterion. If they cannot solve the puzzle within 2 minutes, they finish the puzzle without solving it. Participants are allowed to reverse the paired participant's move.[11] Reversing the partner's move is what I call correction in this paper. Each participant's contribution to a given puzzle is measured by net good moves. Figure 1 shows a sample puzzle screen where a participant is paired with another participant called Valeria and making their move.[12] The paired participant's first name is displayed on the computer screen throughout the puzzle, and when participants select their collaborator to subtly inform the paired participant's gender.

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

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

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

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

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

Part 3: Group work stage

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

Post-experiment

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

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

2.2 Implementation

The experiment was programmed with oTree (Chen, Schonger, and Wickens 2016) and conducted in Italian during November-December 2020. I recruited 464 participants (244 female and 220 male) registered on the Bologna Laboratory for Experiments in Social Science's ORSEE (Greiner 2015) who (i) were students, (ii) were born in Italy, and (iii) had not participated in gender-related experiments before (as far as I could check).[14] The first two conditions were to reduce noise coming from differences in socio-demographic backgrounds and race or/and ethnicity that may be inferred from participants' first name or/and voice, and the last condition was to reduce experimenter demand effects.[15] The number of participants was determined by a power simulation in the pre-analysis plan to achieve 80% power.[16] The experiment is pre-registered with the OSF.[17]

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€, all 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,

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

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

15. Despite that I recruited only Italy-born people, 1 male participant answered in the post-questionnaire that he was from abroad. I include this participant in the analysis anyway but the results are robust to excluding this participant from the data.

16. 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.

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

female participants are more likely to major in humanities, and male participants are more likely to major in natural sciences and engineering, a tendency observed in most OECD countries (see, for example, Carrell, Page, and West 2010).[18] 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.[19]

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.

3 Data description

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

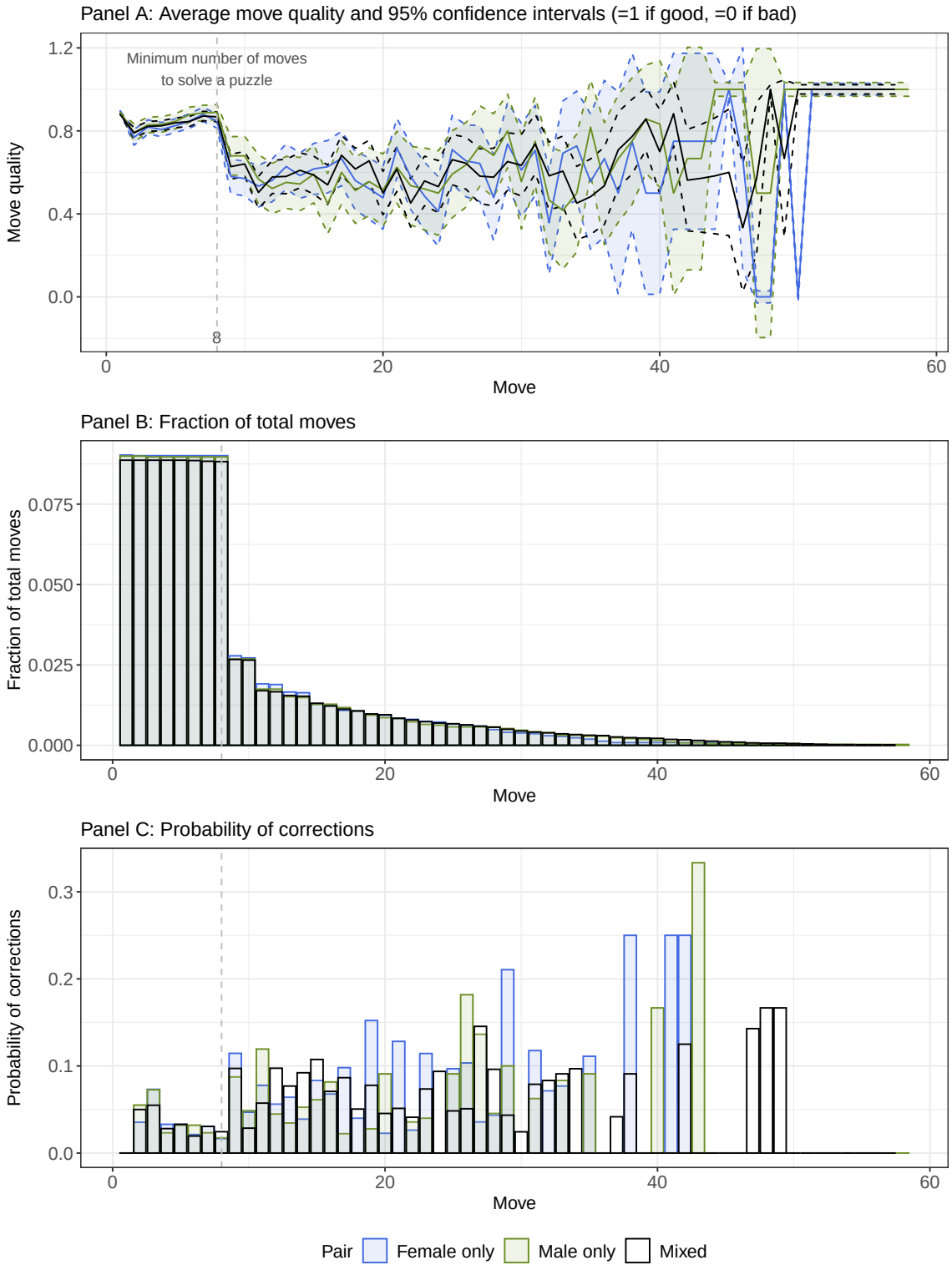
3.1 Move-level data

Figure 3 shows average move quality across moves along with 95% confidence bands (Panel A), fraction of total moves (Panel B), and probability that a correction is happening (Panel C), for female only pairs (blue), male only pairs (green), and mixed gender pairs (black-white). Panel

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

19. No economics PhD student participated in the experiment.

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



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

A shows no statistically significant differences in move quality by own gender or the gender of the partner. Panel B shows that about 71% of the puzzles are solved within a minimum number of moves (the minimum number of moves is 8) and shows that own gender or the gender of the partner does not matter in how fast participants solve the puzzle. Panel C shows that corrections happen across the moves, but there are no systematic differences in the probability that correction is happening by own gender or the gender of the partner.

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: both contribution in part 2 and the number of puzzles solved in part 1, the difference between female and male participants are statistically insignificant at 5% and quantitatively insignificant.[20][21] 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 difference I would find is unlikely to come from their ability difference. Panel A also shows that there are no gender differences in propensity to correct partners, unlike Isaksson (2018), who finds that men correct their partner more often than women, although their result is from move-level data. Finally, the last row of Panel A shows that male participants are slightly more likely to face female partners, although only three percentage points more.

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

Panel B shows that puzzle-solving ability as well as propensity to make corrections (both of a mistake and a right move) of partners paired with female and male participants is the same, suggesting random pairing was successful and that any gender differences I would find are not coming from partners of either gender correct more often. Participants are corrected by their partner in 15-16% of the total puzzles, of which 12-13% are good corrections, and 5-6% are bad corrections, and there are no gender differences in propensity to be corrected.[22]

Panel C shows that participants state they want to collaborate with the partner 71-72% of the time. Participants spend on average 43-44 seconds for each puzzle (the maximum time a pair can spend is 120 seconds) and take 11 moves. 85-86% of the puzzles are solved, and participants and the partner correct each other's move consecutively in 4% of the puzzles.[23] There is no gender difference

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

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

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

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

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).

in any of these outcomes, suggesting any gender differences cannot be attributed to the imbalance in these outcomes.[24]

3.3 Across-round balance

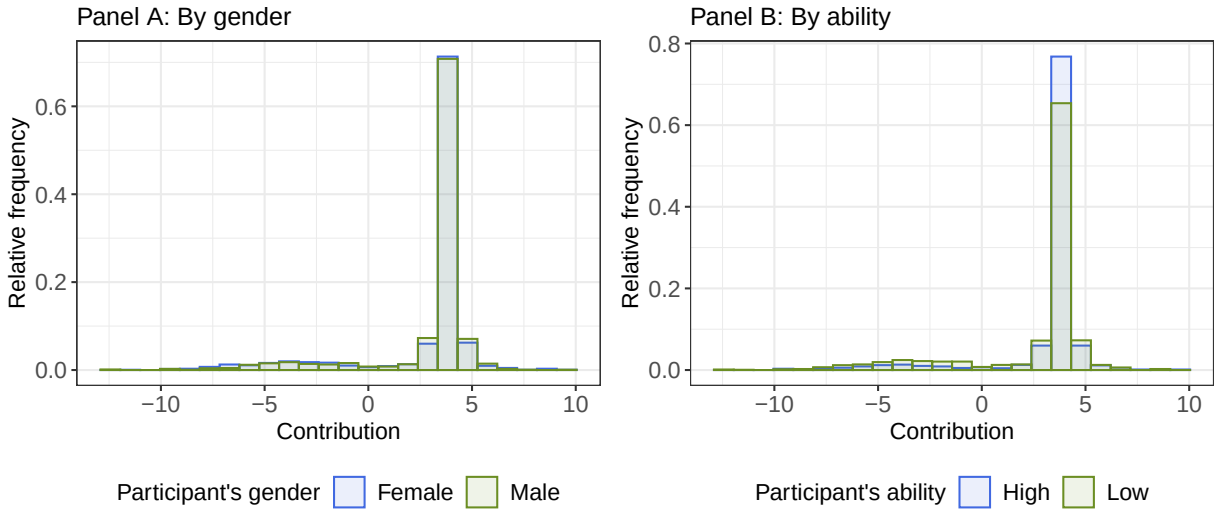
Figure 5 plots average partner gender balance (fraction of female partners, panel A) and puzzle outcomes (panels B-H) across seven rounds along with their 95% confidence intervals (relative to round 1), separately for female (blue) and male participants (green).

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 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

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

Figure 4: Distribution of contribution



Notes: This figure shows the distribution of individual contribution by gender (panel A) and ability (panel B) and shows that most participants contributed to the same degree. Panel A further shows no gender difference in contribution, and panel B further shows that among high-ability people, 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).

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 they 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.

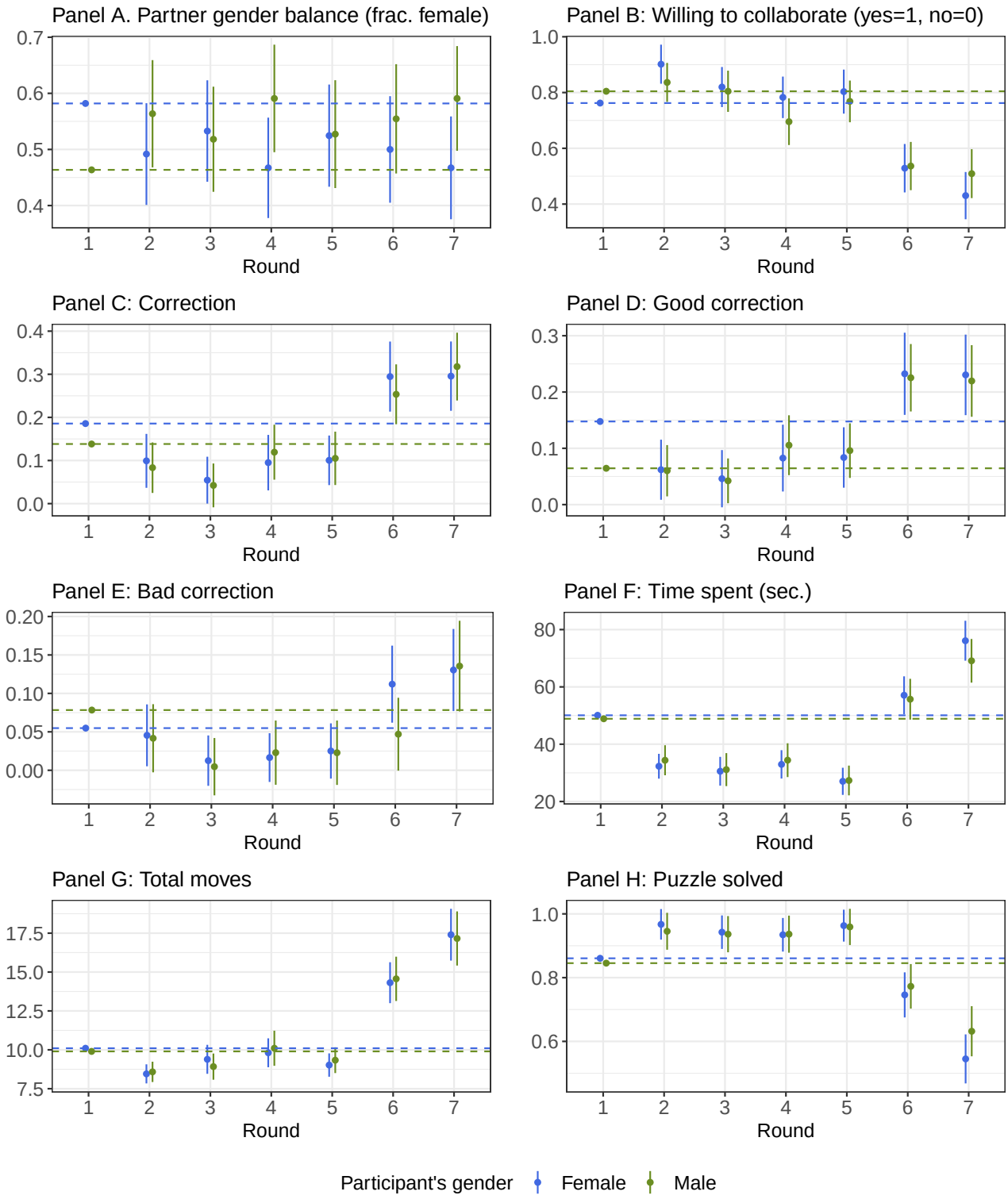
4 Theoretical framework

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 to 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 willing to collaborate with j as 0 for each round t.

The payoff is increasing in i's belief about j's ability. I assume i can partially observe j's move

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.

quality, so i's belief about j's ability is increasing in j's ability 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)

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 is 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 correction and gender convey information about j's ability even if i is fully rational.[25]

First, keeping j's ability 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, hence corrections, regardless of θ. Thus, we have the following proposition:

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

Proposition 2. *A rational agent i with high puzzle-solving ability is more willing to collaborate with j when j made a good correction and 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 correction. On the other hand, if i believes women is worse at the puzzle, then they would consider a correction from a woman as a signal of low ability relative to men's correction. If i believes women and men are equally good at the puzzle, then a correction from a woman or man is irrelevant. Thus, we have the following proposition:

Proposition 3. *A rational agent i's willingness to collaborate with j when j was a woman and made a correction relative to when j was a man and made a correction depends on their 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:*

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

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

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

I consider deviations from these propositions are 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 far away from the solution, 0 otherwise.
- $Female_j \in \{0,1\}$: an indicator variable equals 1 if j is female, 0 otherwise.
- $Contribution_j \in \mathbb{Z}$: js contribution to a puzzle played with i.
- ϵ_{ij}: omitted factors that affect i's likelihood to select j as their collaborator.

and $\mu_i \equiv \sum_{k=1}^{N} \mu^k \mathbb{1}[i=k]$ is individual fixed effects, where N is the total number of participants in the sample and $\mathbb{1}$ is the indicator variable. Standard errors are clustered at the individual level.[26]

More specifically, by 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 paired participant's ability can affect the participant's willingness to collaborate; (ii) there is an effect similar to the reflection effect: participant's personality – for example, meanness – affects their puzzle behavior, which in turn affects the paired participant's behavior. To address the latter point, I add individual fixed effects. To address the former point, I assume that $Contribution_j$ fully captures j's ability perceived by i through j's puzzle moves (not true ability). This assumption is reasonable if we think participants' willingness to collaborate is increasing in the partner's contribution to the puzzle, which is consistent with the fact that participants can partially observe their partners' ability.

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 contribution; 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 the above-median number of puzzles in part 1 in a session they have participated, 0 otherwise.

Other variables are as defined in equation 6.

26. 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.

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.

6 Results

6.1 Response to corrections

Table 3: Response to corrections

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

Notes: This table presents the regression results of equation 6. Columns 1-4 include all participants' willingness to collaborate, but column 1 excludes the partner's contribution and individual fixed effects and column 2 partner's contribution. Column 3 combines good and bad correction as a single dummy variable. Columns 5-7 present the corresponding results for women and columns 8-10 for men. Baseline mean and standard deviation are 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 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 and column 2 partner's contribution. Column 3 combines good and bad correction as a single dummy variable. Columns 5-7 present the corresponding results for women and columns 8-10 for men.

Column 1 shows that when we do not control for between-participants variation, the coefficient

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

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

The coefficient estimate on any correction in column 3 is negative and quantitatively and statistically highly significant and is -0.198 (p-value < 0.01). This suggests that people are 19.8% less willing to collaborate with those who made a correction(s). To offset this effect, a partner's contribution has to increase by 0.79 standard deviations.[27] The corresponding coefficient estimate for women is -0.237 (column 6, p-value < 0.01) and -0.152 for men (column 9, p-value < 0.01). Thus, participants are less willing to collaborate with a person who corrected their move.

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

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

These behaviors are inefficient. They also seem to indicate deviation from the rational agent's benchmark in Proposition 1. However, response to corrections depends on the belief about people's own ability relative to partners and people are in general overconfident, albeit that men are more overconfident (Croson and Gneezy 2009). Thus, these behaviors may not be irrational.

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

6.2 Heterogeneity by participants' ability

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. 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 4 shows 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 Table 3, columns 1-2 include all participants' willingness to collaborate. Columns 3-4 the corresponding results for women and columns 5-6 for men.

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

It is not a problem if this result is coming from high-ability people responding less negatively or even positively to good corrections and more negatively to bad corrections. However, this is not

the case: in column 2, the coefficient estimate on the interaction between good correction and high ability is negative (p-value < 0.05). This effect comes from both women and men, with the effect on men being stronger: the corresponding coefficient estimate for women (column 4) is negative, albeit less so and statistically insignificant and is more negative and statistically significant (p-value $<$ 0.05) for women (in column 6).

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

Thus, even high-ability 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.

6.3 Heterogeneity by partners' gender

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

Looking at column 1, the coefficient estimate on the interaction between 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 suggest that people – both women and men – do not underestimate women's contribution when selecting a collaborator. In other words, people correctly believe that women and men are equally good at solving 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: looking at column 6, the corresponding coefficient estimate for men is -0.119 and marginally significant (p-value < 0.10), while for women it is 0.035 although statistically insignificant (column 4). On the other hand, the positive coefficient estimate on the interaction between female partner and bad correction comes from both women and men: the corresponding coefficient estimate is 0.090 for women (column 4) and 0.168 for men (column 6), although neither of them is statistically significant. Together with the evidence that men believe women are equally

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

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

Notes: This table presents the regression results of equation 8. Columns 1-2 include all participants' willingness to collaborate. Columns 3-4 present the corresponding results for women and columns 5-6 for men. Baseline mean and standard deviation are 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%.

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 being led by women – but they are okay that women make mistakes. As referred to in the introduction, several studies document men's aversion to be 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 a

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, they 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 a paired person.

7.4 Robustness results

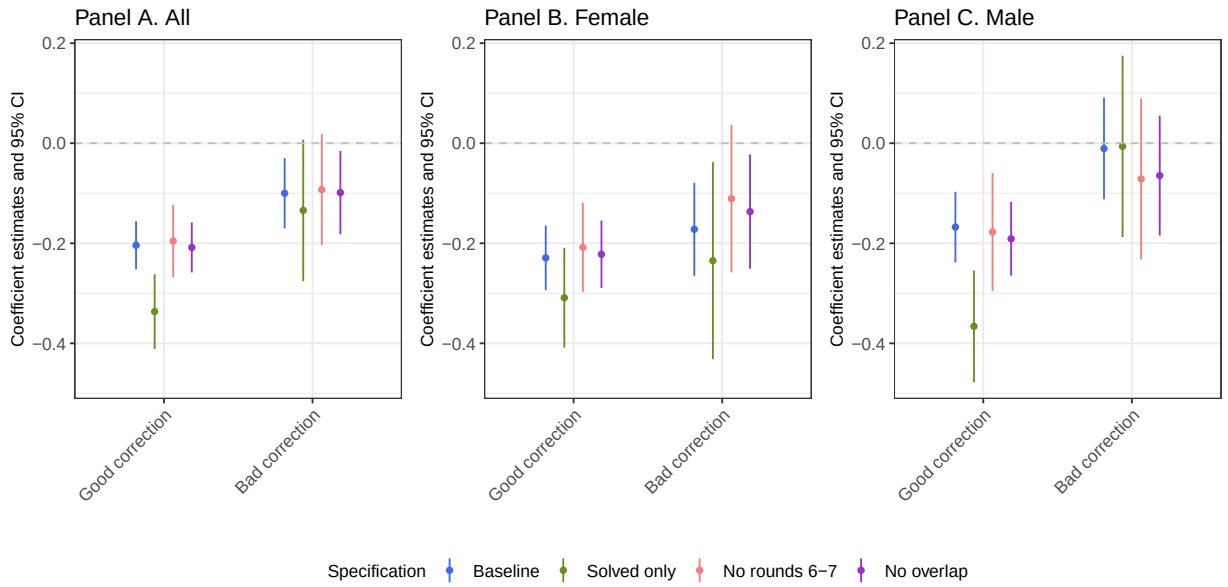
To address these concerns, I re-estimate equations 6, 7, and 8, and plot the coefficient estimates and 95% confidence intervals of the main coefficients of interest in Figures 6, 7, and 8, respectively, with solved puzzles only (green dots and lines), with rounds 1-5 only (red dots and lines), and with 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 interaction 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 interaction 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. Again, the estimates with solved puzzles only present somewhat different evidence; in particular, response to good corrections by female partners is negative although statistically insignificant for women and positive for men. However, both estimates are very close to 0 and do not contradict that the evidence that men react more negatively to women's good correction is only suggestive.

Figure 6: Response to corrections: Robustness



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

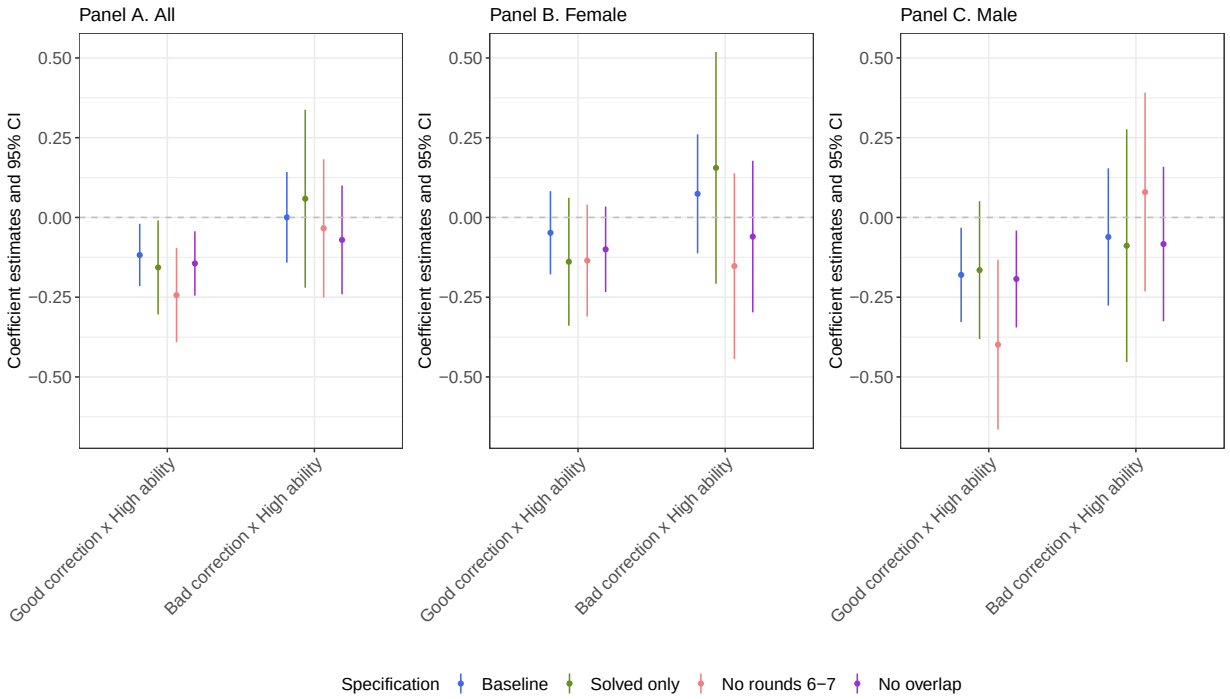
8 Conclusion

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 find suggestive evidence that men respond more negatively to women's corrections that improves group performance, presumably because men do not like to be corrected for their mistakes by women. Thus, dislike to be corrected distorts the optimal selection of talents and penalizes those who correct others' mistakes, and the distortion may be stronger when women correct men.

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 the ability relevant for the participants' work or study – it is not something people have been devoting much of their time to, such as university exams, academic research, or corporate investment projects. Third, participants are equal in my experiment; in the real world, there are sometimes senior-junior relationships, and corrections by junior people may induce stronger negative reactions. Thus, introducing reputation 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

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



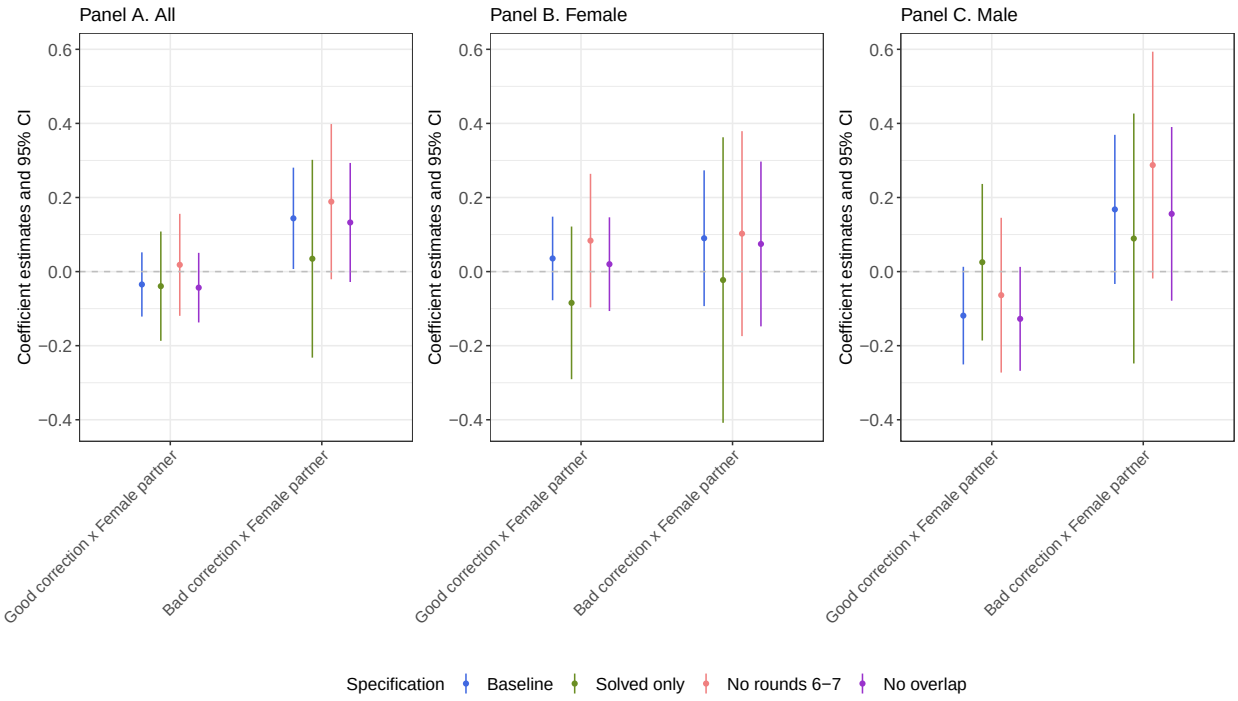
Notes: This figure plots the coefficient estimates and 95% confidence intervals of columns 2, 4, and 6 of Table 4 with solved puzzles only (the green dots and lines), with rounds 1-5 only (the red dots and lines), and with puzzles where only good or bad corrections occurred (the purple dots and lines).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

repeated interactions would mitigate people's negative response to corrections (but they may also magnify the negative response due to rivalry, failure to build a good rapport, etc.). The second limitation is that most participants are bachelor'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 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 results are consistent with self-image concerns and information avoidance (Golman, Hagmann, and Loewenstein 2017).[28] For example, Kszegi (2006) finds that people avoid a difficult task when it reveals their ability. Corroborating this, Castagnetti and Schmacker (2021) find people select information less informative about their ability, and Ewers and Zimmermann (2015) find people exaggerate their ability when others observe it even at the cost of reducing their payoff. Regarding

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

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



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

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.[29]

29. Which 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. 2021. *Improving Workplace Climate in Large Corporations: A Clustered Randomized Intervention.* Working Paper.

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.

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

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

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

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

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

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

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

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

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.*

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.*

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.

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.

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

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

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