Corrections and Gender in Team Collaboration

Yuki Takahashi[*]

November 8, 2024

Abstract

In a quasi-laboratory experiment, I show that individuals are less willing to collaborate with those who corrected them, even when the correction benefited the team. The likely mechanism is negative feedback aversion: more confident individuals are much less willing to collaborate with those who corrected their mistakes but not those who corrected their right actions. Additionally, I find suggestive evidence that men, but not women, are less willing to collaborate with women who corrected their mistakes, potentially due to (inaccurate) beliefs about women's abilities. This reluctance to collaborate with those who corrected them can undermine teamwork, especially in mixed-gender teams.

JEL Classification: M54, D91, J16, C92

Keywords: Correction, Collaboration, Teamwork, Gender, Quasi-laboratory experiment

[*] Department of Economics, Tilburg University. Email: y.takahashi@uvt.nl. This paper was previously circulated as "*Gender Differences in the Cost of Corrections in Group Work*." I am grateful to Maria Bigoni, Siri Isaksson, Boon Han Koh, and Bertil Tungodden, whose feedback was essential for this project, as well as to the participants of the experiments. This paper also greatly benefited from comments at the Discrimination and Diversity Workshop, the ESA Job Market Seminar, the Stanford Institute for Theoretical Economics Conference, the University of Bologna's internal seminar, and the Warwick Economics PhD Conference, among others. The puzzle code used in the experiment is based on Christian König genannt Kersting's code. Lorenzo Golinelli provided excellent technical and administrative assistance. The main experiment was approved by the Bioethics Committee of the University of Bologna (protocol number: 262643), and the follow-up experiment was approved by the Institutional Review Board of the Tilburg School of Economics and Management (approval number: TiSEM_RP1682). The main experimental design and analysis of men's differential responses to women's corrections (columns 3 and 4 of Table 4) and its heterogeneity by men's gender bias (column 3 of Table A2) were preregistered in the OSF registry: https://osf.io/tgyc5.

1 Introduction

Teamwork is essential in most workplaces, and successful collaboration is key to its success. In the corporate sector, more than 50% of workers report their jobs rely on teamwork (Boskamp 2023). In academia, the average number of authors per research article is 2.7 to 5.8 in natural sciences and 2.3 to 3.3 in social sciences (Thelwall and Maflahi 2022). Economics discipline is not the exception: 74% of research articles are written by two or more authors (Jones 2021). However, teamwork involves human interactions, and several factors can inhibit its efficient functioning. For instance, workers are hesitant to seek advice from senior colleagues due to concerns about signaling weakness in their abilities (Chandrasekhar, Golub, and Yang 2019). Additionally, a non-supportive atmosphere reduces mutual reciprocation among workers (Alan, Corekcioglu, and Sutter 2023). Further, gender stereotyping and self-stereotyping can hinder teamwork: workers are less likely to contribute their ideas in gender-incongruent domains (Coffman 2014), and even if they do, colleagues may be less likely to use them (Coffman, Flikkema, and Shurchkov 2021).

Another potential obstacle to efficient teamwork is the correction of colleagues' mistakes. For example, a worker might need to correct a colleague's miscalculated numbers in a presentation slide or a flawed conclusion in a report. While these corrections are essential for team success, they can damage relationships if taken personally. Additionally, women are less likely than men to correct others in academia (Klinowski 2023), potentially due to men's aversion to being corrected by women, as suggested by anecdotal evidence (e.g., Cooper 2018).[1] If these are the case, corrections may create friction, particularly in mixed-gender teams.

This paper investigates whether individuals are less willing to collaborate with those who corrected them and whether men are particularly less willing to do so. I define collaboration as working with others toward the same goal and correction as overriding what others have done. To answer these questions, I design a quasi-laboratory experiment – a hybrid of laboratory and online experiments – where group formation is randomized. The experiment allows participants to correct each other and express their willingness to collaborate without fear of external consequences, such as interpersonal frictions outside the experiment, which are difficult with observational data or in a field setting. Participants are grouped into teams of eight and perform a collaborative task in pairs seven times, each time with a different partner. Each time participants complete the task, they privately indicate whether they would prefer to collaborate with their current partner for the final stage, which is the main source of earnings (up to 20€ in 12 minutes), providing a strong incentive to select a capable partner.

For the team task, I use the number-sliding puzzle from Isaksson (2018), which allows for an objective measurement of each participant's contribution and the classification of moves as either good (advancing the puzzle) or bad (hindering progress). The task also provides a clear definition of a correction – reversing a partner's move – making it comparable across participants. At the

[1] A quote from Sarah Cooper's book goes as follows: "As women, we might be tempted to say, 'Excuse me, you made a small error here.' DO NOT SAY THIS. As non-threatening women, we must avoid that instinct because it serves no one, least of all ourselves."

beginning of the experiment, participants are informed about the notion of good and bad moves, how to solve the puzzles efficiently, and how the collaborator will be selected for the final stage. To avoid concerns about backlash, the experiment is one-shot, and participants remain largely anonymous.

I first confirm that the participants understand good and bad moves, as they are more likely to select those who contributed more as collaborators. I also find that men and women contribute equally to the puzzle, and in the absence of corrections, participants are equally likely to select male and female collaborators with comparable contributions. However, after controlling for contributions, participants are less willing to collaborate with those who corrected them, even when the corrections are beneficial to the team. This effect is substantial, reducing the likelihood of collaboration by about 20 percentage points or approximately 25% relative to the baseline mean, which would require an additional contribution of 0.79 standard deviations to compensate.

The likely mechanism behind these behaviors is negative feedback aversion. Consistent with the literature on ego and information processing (Kőszegi 2006; Eil and Rao 2011), participants who are more confident in their abilities are significantly less willing to collaborate with those who corrected their mistakes, but not their right moves.

I also find suggestive evidence that men, but not women, are less willing to collaborate with women than with equivalent men who corrected their mistakes. This may be driven by (inaccurate) beliefs about women's puzzle-solving abilities: men tend to view the puzzle as slightly male-typed, while women perceive it as gender-neutral. Interestingly, men show no such aversion when women correct their right moves. Taken together, the unwillingness to collaborate with those who corrected them can indeed be another obstacle to efficient teamwork, especially in mixed-gender teams.

The main contribution of this paper is to demonstrate that individuals' reluctance to collaborate with those who corrected them can be a factor hindering teamwork, particularly in gender-mixed teams, which adds to the literature on factors that prevent successful teamwork. The literature shows that individuals are less willing to seek advice from senior colleagues due to concerns about signaling weakness in their abilities (Chandrasekhar, Golub, and Yang 2019). Additionally, they are less likely to reciprocate in non-supportive work environments (Alan, Corekcioglu, and Sutter 2023). Gender stereotyping and self-stereotyping also play a role: individuals are less willing to contribute ideas in gender-incongruent domains due to self-stereotyping (Coffman 2014). Even when they do contribute, their ideas are less likely to be utilized by teams (Coffman, Flikkema, and Shurchkov 2021). Furthermore, men tend to dominate team discussions, even when their abilities are lower than those of their female colleagues, which ultimately reduces team performance (Hardt, Mayer, and Rincke 2024).

This paper also adds to the literature on gender disparities in teamwork. The closest study to mine is Guo and Recalde (2023), who finds that individuals are more likely to overwrite women's opinions than men's in a male-typed task. I show the other side of the coin: men are more averse to being corrected by women than by men in a slightly male-typed task, complementing Guo and Recalde. Similarly, Isaksson (2018), using the same puzzle, finds that women claim less credit for their contributions than men, especially in difficult puzzles, and that men correct their partners more often than women. Corroborating these findings, Klinowski (2023) shows that female scientists are less likely than their male counterparts to criticize or correct other scientists' publications.

On team leadership, Born, Ranehill, and Sandberg (2022) find that women are less likely to be selected as leaders, although they perform equally well once selected (Heursen, Ranehill, and Weber 2022). Finally, Abel (2024) finds that workers become less motivated when criticized by female managers but not when they receive praise from them. Relatedly, Sinclair and Kunda (2000) find that college students become more gender-biased when female instructors give them low grades but not when they give high grades. Consistent with these findings, I observe asymmetric responses to women's positive and negative corrections.

The remainder of the paper is structured as follows. Section 2 details the design, procedure, and implementation of the experiment. Section 3 describes the data obtained from the experiment. Section 4 outlines the empirical strategy, followed by the analysis of the effects of receiving corrections in Section 5, and the analysis of gender-specific responses to corrections in Section 6. Section 7 assesses the robustness of the results. Finally, Section 8 summarizes the findings and discusses their implications for teamwork and gender dynamics in collaborative environments.

2 Experiment

This study consists of two experiments: the main experiment and a follow-up experiment. The main experiment collects data on collaborator preferences and other key variables, while the follow-up experiment gathers data on individuals' perceptions of the genderness of the task used in the main experiment. The experimental instructions for both experiments can be found in the Online Appendix B.

2.1 Main experiment

The main experiment was conducted in a quasi-laboratory format. Participants and experimenters were connected via Zoom throughout the experiment. Participants' cameras and microphones were turned off except at the beginning of the session, and participants completed the tasks remotely using their computers. The experiment followed the protocols of a traditional physical laboratory setting.

The experiment consisted of three parts. In Part 1, participants solved the puzzle individually to familiarize themselves with the task and for me to assess their puzzle-solving abilities. In Part 2, participants learned the rules of Part 3 and stated their collaborator preferences after solving one puzzle with each potential collaborator. In Part 3, participants worked on puzzles with collaborators selected based on their preferences from Part 2. At the beginning of each part, participants answered comprehension questions to ensure they understood the instructions. Figure 1 summarizes the flow of the experiment, which I explain in detail below.

Figure 1: Flowchart of the main experiment

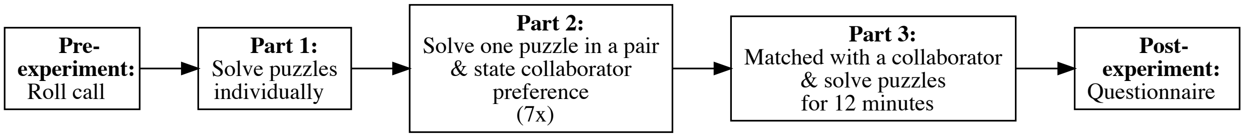
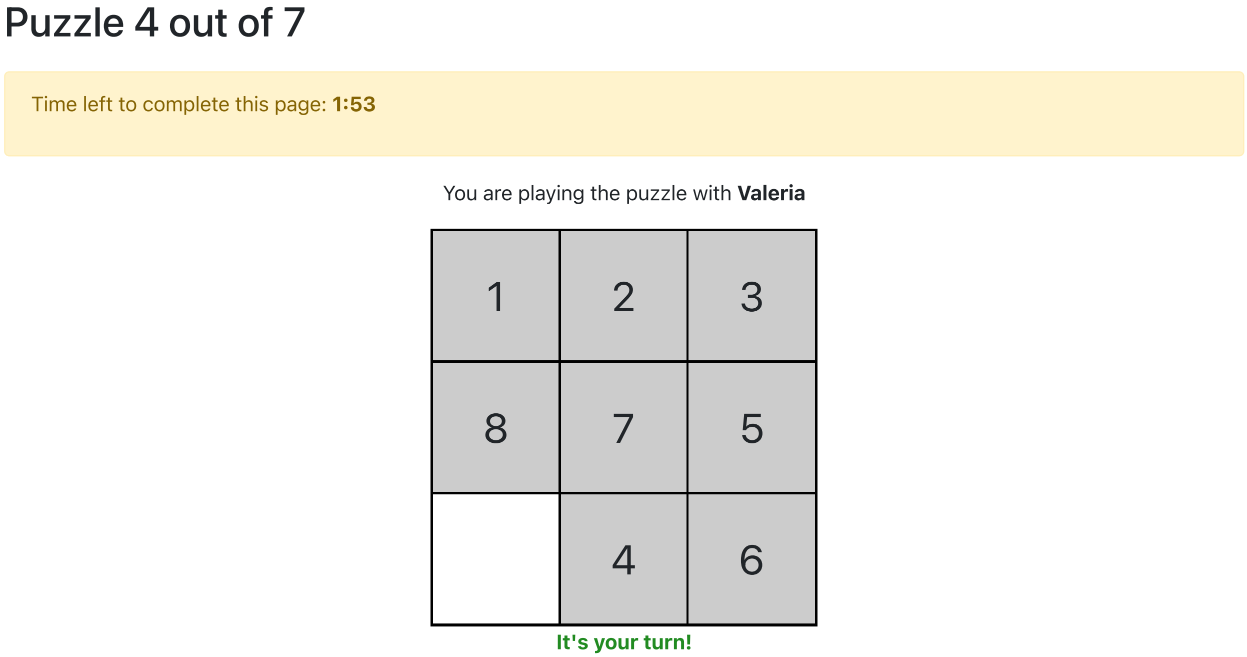

Pre-experiment: Roll call
Part 1: Solve puzzles individually
Part 2: Solve one puzzle in a pair & state collaborator preference (7x)
Part 3: Matched with a collaborator & solve puzzles for 12 minutes
Post-experiment: Questionnaire

Figure 2: Puzzle screen



Puzzle 4 out of 7

Time left to complete this page: **1:53**

You are playing the puzzle with **Valeria**

1	2	3
8	7	5
	4	6

It's your turn!

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 in the experiment are in Italian.

The team task

I used the sliding puzzle developed by Isaksson (2018) as the team task. The puzzle consists of eight numbered tiles arranged in a 3x3 frame, with the goal of placing the tiles in numerical order (see Figure 2 for an example). Participants played in pairs, taking turns to make moves. Each participant was required to make a move on their turn, and passing was not allowed.

This puzzle offers key advantages for measuring contributions in teamwork. Using the Breadth-First Search algorithm, I was able to objectively classify each move as either "good" (bringing the puzzle closer to the solution) or "bad" (moving further from the solution). Participants' individual contributions were measured as their net good moves – the number of good moves minus the number of bad moves they made in a given puzzle.

A correction was defined as reversing a partner's move, which allowed for objective comparison of corrections between participants, since corrections themselves count as moves.[2] This setup captures

[2] Because some corrections happen early in the puzzle and others later, I capture the average effect of a correction in the analysis.

a crucial aspect of teamwork, where participants work towards the same goal, but the quality of individual moves and corrections is only partially observable to the participants (fully observable to the experimenter). This partial observability allows for motivated reasoning, in which participants interpret corrections in a self-serving manner (Kunda 1990; Chance and Norton 2015).

At each stage of the puzzle, there was only one correct strategy: making a good move.[3] Multiple good or bad moves could be present in a given stage, but all were equal in quality. The puzzle had no path dependency, meaning the sequence of previous moves did not affect future moves.

Pre-experiment

Participants entered the Zoom waiting room at their assigned session time. Upon verification of their registration, they received a link to the virtual experiment room and provided their first name, last name, and registration email, which was used to match their earnings with their payment information on the laboratory's subject database.

As participants arrived and verified, they were admitted to the Zoom meeting room individually, and their names were displayed as their first name. If multiple participants had the same first name, a number was appended (e.g., Giovanni2). Because Italian first names have little variation (ISTAT 2024), showing first names is unlikely to reveal participants' identities.[4] A roll call was then conducted to disclose participants' gender without making it explicitly salient (Bordalo et al. 2019; Coffman, Flikkema, and Shurchkov 2021; Erkal, Gangadharan, and Koh 2023). During the roll call, participants responded verbally via microphone, revealing their gender.[5] Figure 3 shows the Zoom screen participants viewed during the roll call.

Afterward, I read the experimental instructions and answered participants' questions. During the experiment, participants communicated with the experimenter via Zoom's private chat.

Part 1: Individual practice stage

In Part 1, participants were given in-depth instructions on how to efficiently solve the puzzle (minimizing total moves) and were asked to complete comprehension questions to confirm their understanding.[6] Participants then worked on the puzzle individually for 4 minutes, solving as many puzzles as they could (maximum 15 puzzles), with puzzles increasing in difficulty. Each correct solution was incentivized with 0.2€. After 4 minutes are up, they receive information on how many puzzles they have solved. This part familiarized participants with the puzzle and provided a measure of their ability.

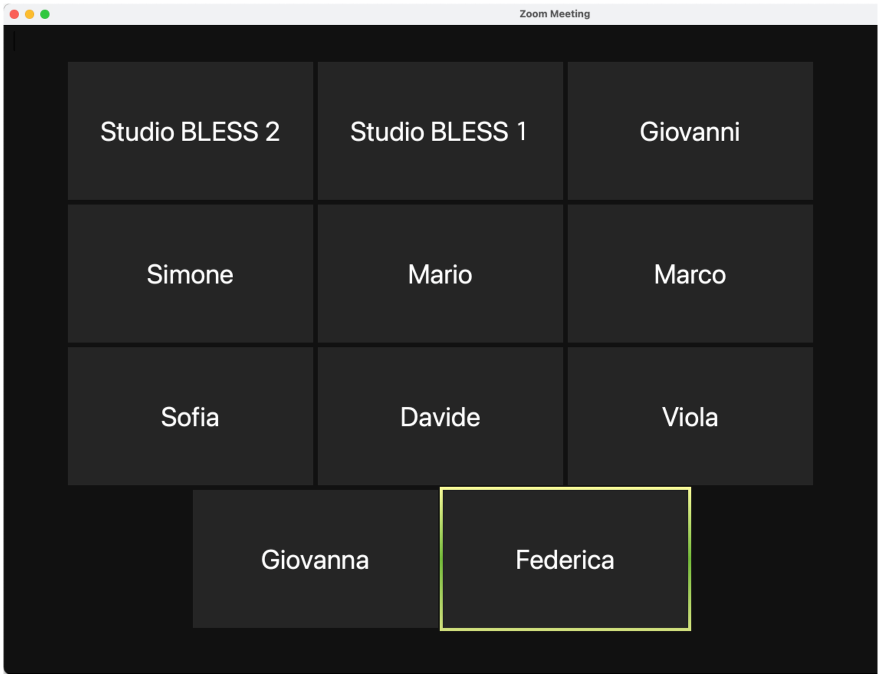
[3] This assumes that both players are trying to solve the puzzle; I show in Figure 8 that the results are robust to the exclusion of puzzles where either player might not be trying to solve the puzzle.

[4] For children born in 1999, the earliest available year and the closest year to my participants' years of birth, the top 10 names cover 25.5% of girls' and 29.6% of boys' names.

[5] Participants' response was kept short. For example, the most common responses were "sì" (yes), "presente" (I am present), "io" (me), and "ci sono" (I am here).

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

Figure 3: Zoom screen



Notes: This figure shows the Zoom screen participants would see during the roll call.

Part 2: Collaborator selection stage

Part 2 consisted of seven rounds. Before starting, participants were instructed on the rules of Part 3. This part was modeled after the speed dating experiments by Fisman et al. (2006, 2008). Participants were divided into groups of eight based on their abilities as measured in Part 1, to minimize ability differences and make corrections and gender more salient.

In each round, participants were randomly paired with another member of their group and worked together to solve a puzzle by alternating their moves. The first mover was randomly chosen, and both participants were aware of this selection criterion. If a puzzle was not solved within 2 minutes, the round ended. Participants were allowed to correct their partner's moves.[7] After each puzzle, participants privately stated whether they would like to collaborate with their current partner in Part 3.[8] The pairing was conducted using a perfect stranger matching procedure, ensuring each participant was paired with every other member of their group exactly once. Figure 2 shows a sample puzzle screen in which 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.

[7] Solving the puzzle itself is not incentivized, so participants who do not want to collaborate with a given partner or fear receiving 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.

[8] The sequence of puzzles is the same for all pairs in all sessions. The puzzle difficulty is kept the same across all seven rounds. Based on a pilot, I set the minimum number of moves to solve the puzzles to be eight so that the puzzles are neither too easy nor too difficult to solve.

At the end of Part 2, participants were matched for Part 3 based on mutual collaboration preferences using an algorithm adapted from Fisman et al. (2006, 2008). This matching algorithm is incentive compatible under the assumption that payoff is the primary concern for the collaborator selection. While other factors matter in real life, they would not play an important role in such a short experiment.[9] The matching process was explained in detail at the beginning of Part 2 to ensure participants understood the implications of their preferences.

Part 3: Teamwork stage

In Part 3, participants worked in their assigned pairs for 12 minutes, alternating moves to solve puzzles. Each correct solution earned the pair 1€. The first mover was randomly determined at the start of each puzzle, and participants could solve up to 20 puzzles, with difficulty increasing as the game progressed.

Post-experiment

After completing the puzzles, participants answered a short questionnaire. This included (i) six questions on hostile and benevolent sexism, as used by Karpowitz et al. (2024), to measure participants' gender biases, and (ii) questions on demographics and their impressions of the experiment. The questionnaire helped determine whether participants anticipated that the experiment was related to gender. No evidence suggested that participants were aware of the gender focus.

Earnings were calculated and communicated to participants privately. They later received their earnings via PayPal.

Implementation and participant characteristics

The experiment was programmed using oTree (Chen, Schonger, and Wickens 2016) and conducted in Italian in November and December 2020. A total of 464 participants (220 male, 244 female) were recruited from the Bologna Laboratory for Experiments in Social Science's ORSEE database (Greiner 2015). The participant pool was restricted to students born in Italy who had not previously participated in gender-related experiments.[10,11] The first two conditions were imposed to reduce variability in socio-demographic backgrounds and to control for the influence of race or ethnicity that could be inferred from participants' names or voices.[12]

[9] Specifically, the matching was done as follows: (i) for every participant i, I counted the number of matches; that is, the number of other participants in the group who were willing to collaborate with i and with whom i was willing to collaborate in part 3. (ii) I randomly chose one participant. If the chosen participant had only one match, I paired them up and let them work together in part 3. If the chosen participant has more than one match, I randomly chose one of the matches. (iii) I excluded participants who had been paired and repeated (i)-(ii) until no feasible match was left. (iv) If some participants were left unpaired, I paired them up randomly.

[10] I include 16 participants from a pilot session where the experimental instructions were slightly different. The results are robust to the exclusion of these 16 participants.

[11] The laboratory prohibits deception, so no participant participated in an experiment with deception.

[12] Despite only recruiting individuals born in Italy, one male participant answered in the post-questionnaire that he was born abroad. I included this participant in the analysis anyway but the results are robust to excluding this participant.

Twenty-nine sessions were conducted with 16 participants each, and the average session lasted 70 minutes. Participants earned an average of 11.55€, with a maximum of 25€ and a minimum of 2€, including the 2€ show-up fee.

Online Appendix Table A1 provides a summary of participants' characteristics. Male participants were slightly older (by 1.41 years) and exhibited slightly higher gender bias (by 0.12 points) than female participants. In addition, male participants are more likely to major in natural sciences and engineering and less likely to major in humanities, a tendency observed in most OECD countries (see, for example, Carrell, Page, and West 2010).[13] Most participants were either bachelor's or master's students, with very few PhD students. No economics PhD students participated in the experiment.

2.2 Follow-up experiment

The follow-up experiment was designed to collect data on participants' perceptions of the genderness of the puzzle used in the main experiment. This experiment had two parts. In Part 1, participants solved one puzzle individually to become familiar with the task. In Part 2, they made an incentivized guess about which gender – male or female – solved more puzzles in Part 1 of the main experiment, using a 7-point Likert scale. Participants whose guess was correct earned an additional £1 on top of the completion fee. Afterward, participants answered a short questionnaire on demographics.

The follow-up experiment was also programmed using oTree and conducted in Italian in August 2024. A total of 80 participants (40 male, 40 female) were recruited via Prolific. To ensure similarity to the original participant pool, participants were restricted to students living in Italy with Italian nationality and Italian as their first language. The experiment lasted approximately 4 minutes on average, and participants earned an average of £1.86, including the completion fee of £1.5.

3 Data

I use data from Part 2 of the experiment, where we observe collaborator selection decisions. I aggregate move-level data for each puzzle to link puzzle behavior with collaborator selection decisions.[14]

3.1 Data description

Table 1 provides a summary of participants' own puzzle behaviors (Panel A), their partner's behaviors (Panel B), and puzzle outcomes (Panel C). Panel A shows no significant gender differences in puzzle-solving ability: for both contributions in Part 2 and the number of puzzles solved in Part 1, the

[13] Individual fixed effects in the analysis control for participants' major.

[14] Online Appendix Figure A1 summarizes the move-level data and shows no statistically significant differences in move quality by participants' gender or the gender of their partner. It also shows no systematic differences in the likelihood of making a correction based on one's own gender or that of the partner, and gender does not affect how quickly participants solve the puzzle.

Table 1: Own and partner's puzzle behaviors and puzzle outcomes

	Male		Female		Difference		
	(N=1540)		(N=1708)		(Male – Female)		
	Mean	SD	Mean	SD	Mean	SE	P-value
Panel A: Own behaviors							
Contribution	3.14	2.64	2.98	2.93	0.16	0.10	0.11
# puzzles solved in part 1	8.80	2.34	8.36	2.41	0.44	0.22	0.05
Any correction	0.16	0.36	0.15	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	0.05	0.22	0.06	0.23	0.00	0.01	0.70
(Fraction of female partners)	0.54	0.50	0.51	0.50	0.03	0.02	0.03
Panel B: Partner's behaviors							
Contribution	3.07	2.87	3.04	2.73	0.03	0.10	0.77
# puzzles solved in part 1	8.57	2.43	8.58	2.35	-0.01	0.16	0.93
Any correction	0.15	0.36	0.16	0.37	-0.01	0.01	0.51
Good correction	0.12	0.32	0.13	0.33	-0.01	0.01	0.44
Bad correction	0.05	0.22	0.06	0.23	-0.01	0.01	0.44
Panel C: Puzzle outcomes							
Willing to collaborate (yes=1, no=0)	0.71	0.45	0.72	0.45	-0.01	0.02	0.49
Willing to collaborate (residualized)	0.00	0.42	0.00	0.42	0.00	0.00	0.46
Time spent (second)	42.99	35.76	43.74	36.15	-0.74	1.28	0.56
Total moves	11.21	7.70	11.18	7.46	0.03	0.28	0.92
Puzzle solved	0.86	0.35	0.85	0.36	0.01	0.01	0.43
Consecutive correction	0.04	0.21	0.04	0.20	0.00	0.01	0.81

Notes: This table summarizes participants' own (Panel A) and their partner's puzzle behaviors (Panel B), as well as puzzle outcomes (Panel C). P-values of the gender differences 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).

gender differences are statistically insignificant at the 5% level and quantitatively small.[15,16] These results align with those of Isaksson (2018), who also found no gender differences in contributions or puzzle-solving using the same puzzle. This suggests that any gender differences observed in this study are unlikely to stem from ability differences between male and female participants.

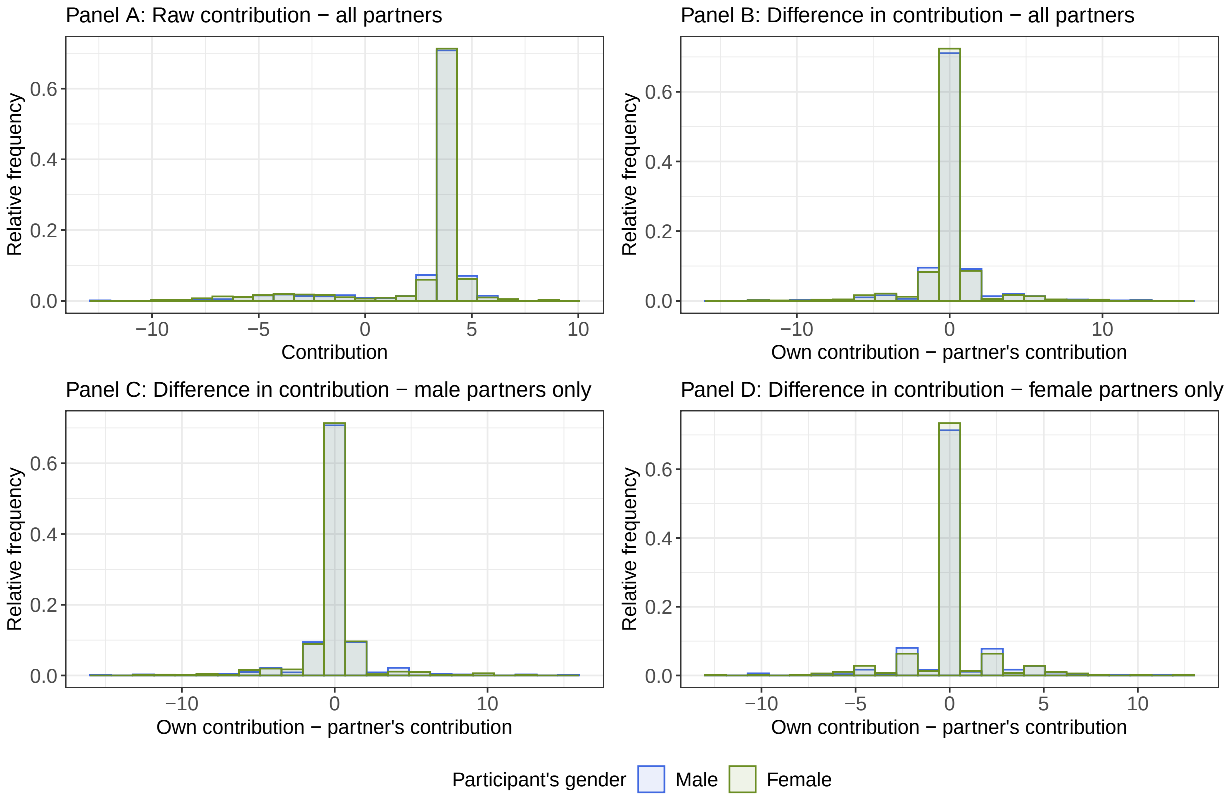
Panel A also shows no gender differences in the likelihood of correcting partners, which contrasts with Isaksson (2018)'s finding that men correct their partners more often, and with Klinowski (2023), who found that men are more likely to point out (and penalize) others' mistakes. Of the 495 puzzles in which at least one correction occurred, 354 (72%) involved only one correction, and 141 (28%) involved more than one. Among the puzzles with more than one correction, 51% experienced only good corrections, 6% experienced only bad corrections, and 43% experienced both. Later, I show that the results remain robust when excluding puzzles with overlapping corrections. Finally, the

[15] The number of puzzles solved in Part 1 is marginally significant, but quantitatively insignificant.

[16] The correlation coefficient between contributions and the number of puzzles solved in Part 1 is 0.1059, with a p-value below 0.001 (standard errors clustered at the individual level).

last row of Panel A shows that male participants are slightly more likely to have female partners, though the difference is small (three percentage points).

Figure 4: Distribution of contributions



Notes: This figure presents the distribution of individual contributions by gender. Panel A shows the raw contribution distribution, while Panels B-D show the difference in contributions between own and partner's moves, broken down by male partners (Panel C) and female partners (Panel D). Contribution is defined as net good moves.

Expanding on Panel A of Table 1, Panel A of Figure 4 shows the distribution of contributions by gender. The figure illustrates that most participants contribute similarly, with men and women performing equally well. In about 70% of the puzzles, participants contributed 4 net good moves (good moves minus bad moves), and in 90% of puzzles, contributions ranged from 3 to 5. Panels B-D further confirm that these patterns hold across different gender pairings. While some outliers exist, I later demonstrate that excluding these outliers does not change the results.

Panel B of Table 1 shows that puzzle-solving ability and the likelihood of correcting partners' moves (both correct and incorrect moves) are consistent across gender pairings, indicating that the random pairing was successful. Participants were corrected in 15-16% of the puzzles, with 12-13% of corrections being good and 5-6% being bad. There are no significant gender differences in the likelihood of being corrected.[17]

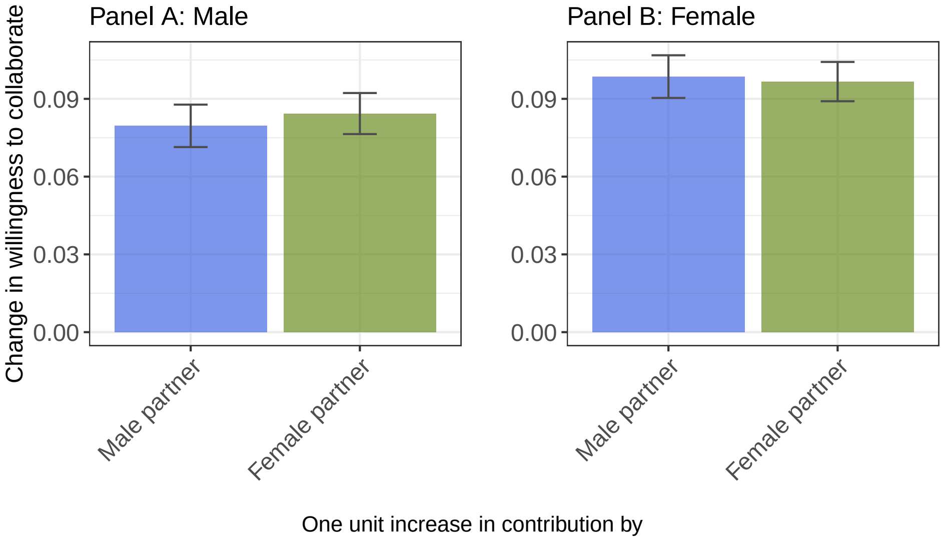
Panel C shows that participants chose to collaborate with their partner 71-72% of the time, with

[17] The sum of good and bad corrections does not equal the total percentage of corrections because some puzzles contain both good and bad corrections. The results are robust to excluding these puzzles, as shown in Figure 8.

sufficient within-subject variation.[18] Participants took an average of 43-44 seconds per puzzle (the maximum allowed time was 120 seconds) and made an average of 11 moves. In 85-86% of the puzzles, participants successfully solved the puzzle, and in 4% of the puzzles, participants corrected their partner's moves consecutively. Notably, in puzzles with consecutive corrections, the likelihood of selecting the partner as a collaborator dropped from 78% to 27%. There were no gender differences in these outcomes, further indicating that gender did not drive any imbalances in these variables. I later show that the results are robust to excluding unsolved puzzles and those where participants spent more than 40 or 60 seconds.

3.2 Response to contribution

Figure 5: Change in willingness to collaborate by one unit increase in partner's contribution



Notes: This figure shows changes in men's (Panel A) and women's (Panel B) willingness to collaborate following a one-unit increase in male (blue) and female (green) partners' contributions, with 95% confidence intervals calculated using standard errors clustered at the individual level.

Both men and women correctly respond to their partners' contribution, regardless of the partners' gender. Figure 5 shows the change in men's (Panel A) and women's (Panel B) willingness to collaborate for each one-unit increase in the contribution of male (blue) and female (green) partners, along with 95% confidence intervals. The figure indicates that both men and women are more likely to collaborate with partners who contribute more, irrespective of gender. I show later that participants do not exhibit a preference for male or female collaborators. The smaller increase in men's willingness to collaborate may be due to their overconfidence, leading them to underestimate their partner's abilities. The positive response to contributions indicates that participants understand the concept of good and bad moves.[19]

[18] This variation is demonstrated by the standard deviation of the residualized Willingness to collaborate. Residuals were obtained by regressing willingness to collaborate on individual fixed effects and extracting the residuals.

[19] As inferred from the positive contribution, the high percentage of solved puzzles, and the fact that participants

3.3 Across-round balance

Finally, key variables remain balanced across rounds, as shown in Figure 6, although some imbalance appears in rounds 6 and 7, where participants are less willing to collaborate, experience more corrections, and are less likely to solve the puzzle. I later show that the results are robust to excluding these rounds.[20]

4 Empirical strategy

I estimate the following equation via OLS to examine how corrections affect a participant's willingness to collaborate with the partner who corrected them:

$$Select_{ij} = \beta_1 CorrectedGood_{ij} + \beta_2 CorrectedBad_{ij} + \beta_3 Female_j + \delta Contribution_j + \mu_i + \epsilon_{ij} \quad (1)$$

where each variable is defined as follows:

- $Select_{ij} \in \{0, 1\}$: an indicator variable equal to 1 if participant i selects j as their collaborator, and 0 otherwise.
- $CorrectedGood_{ij} \in \{0, 1\}$: an indicator variable equal to 1 if participant j corrected i and moved the puzzle closer to the solution, and 0 otherwise.
- $CorrectedBad_{ij} \in \{0, 1\}$: an indicator variable equal to 1 if participant j corrected i but moved the puzzle further from the solution, and 0 otherwise.
- $Female_j \in \{0, 1\}$: an indicator variable equal to 1 if participant j is female, and 0 otherwise.
- $Contribution_j \in \mathbb{Z}$: j's contribution to the puzzle played with i (measured by net good moves).
- ϵ_{ij}: the error term.

$\mu_i \equiv \sum_{k=1}^{N} \mu^k \mathbb{1}[i = k]$ represents the individual fixed effects, where N is the total number of participants, and $\mathbb{1}$ is the indicator variable. Standard errors are clustered at the individual level.[21]

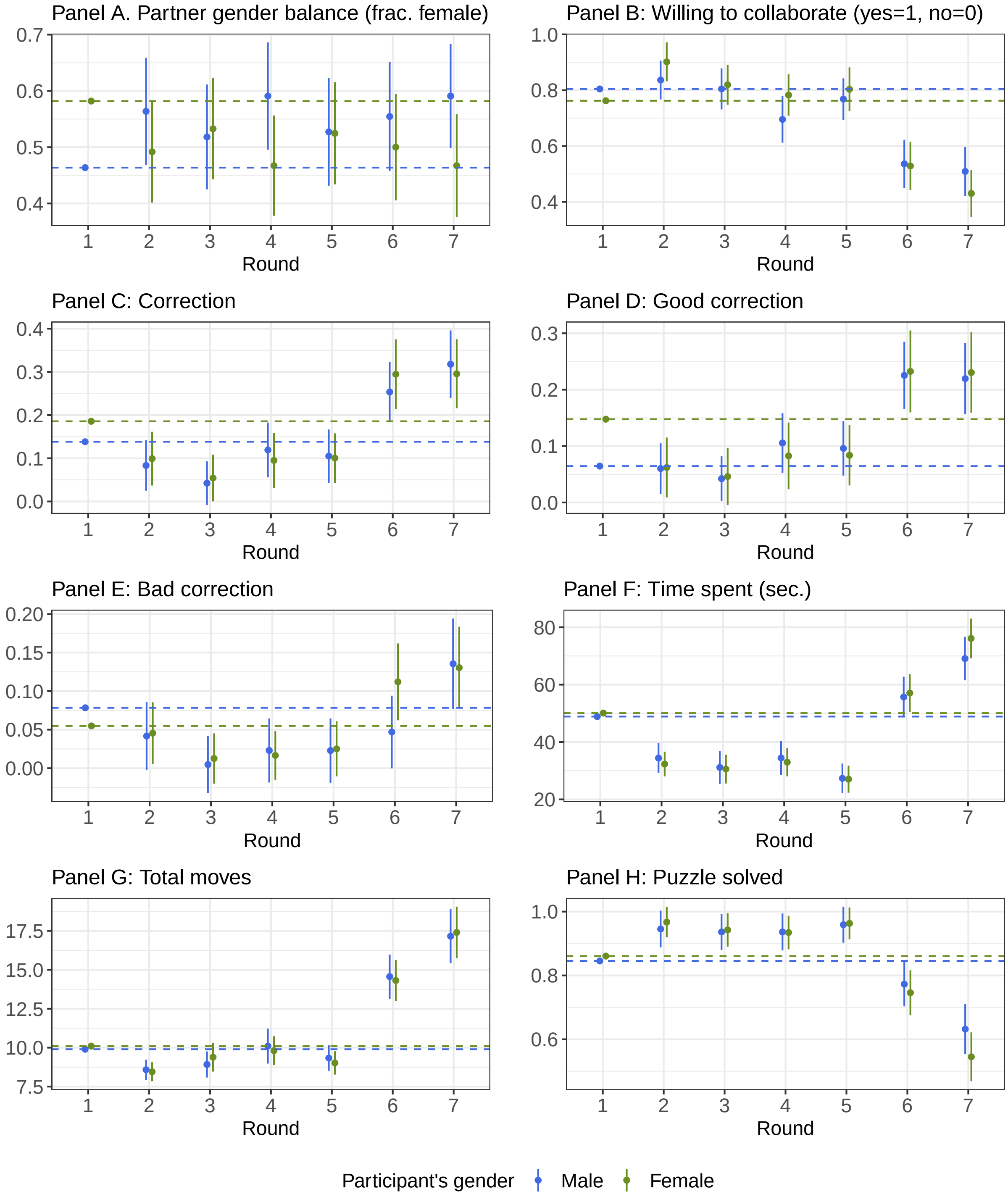
I exploit the random pairing of participants, conditional on individual fixed effects, for causal identification. The random pairing is conditional because the groups of eight individuals in Part 2 are formed based on each participant's performance in Part 1. Specifically, I control for all observable characteristics a participant might consider when assessing their partner, including the partner's gender, whether the partner made a correction, and the partner's perceived puzzle-solving ability. Conditional on these observables, corrections occur due to specific puzzle configurations, which are held constant in terms of objective difficulty across rounds, although some participants may find certain configurations more challenging than others. I later demonstrate that (i) the time taken to

are more willing to collaborate with those who contribute more, the puzzles do not appear too difficult for participants. See Online Appendix Figure A2, which shows the participants' perceived puzzle difficulty from the post-experiment questionnaire.

[20] One imbalance concerns the gender balance of partners in round 1, where male participants are more likely to have female partners. However, this imbalance does not persist in rounds 2-7.

[21] The treatment unit is participant i. Although the same participant appears twice (once as i and once as j), j is passive in the collaborator selection process.

Figure 6: Balance across rounds



Notes: This figure presents point estimates and 95% confidence intervals of βs from an OLS regression of gender balance (female dummy) and puzzle outcomes, separated by male (blue) and female (green) participants: $y_{ij} = \beta_1 + \sum_{k=2}^{7} \beta_k \mathbb{1}[t_{ij} = k] + \epsilon_{ij}$, 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 the estimates of β_2-β_7 to make the figure easier to look at. Standard errors are clustered at the individual level.

solve the puzzle does not mediate the results, and (ii) the results are robust to different functional forms for the partner's perceived puzzle-solving ability.

Coefficients of interest and their interpretations

The coefficients of interest are β_1 and β_2. β_1 captures whether participants' willingness to collaborate with a partner who corrected them and moved the puzzle closer to the solution differs from their willingness to collaborate with a partner who did not correct them, holding perceived ability constant. β_2 captures the same, but when the correction moved the puzzle further away from the solution.

Because perceived ability is controlled for, β_1 and β_2 function as signals of the partner's ability. Assuming participants are rational and can partially observe the quality of each move, β_1 serves as a positive signal about the partner's ability, as it reflects a correction of a bad move, and is expected to be positive. β_2 serves as a negative signal about the partner's ability, as it reflects a correction of a good move, and is expected to be negative.

As participants' ability to observe the quality of moves increases, both β_1 and β_2 should approach zero, as these signals become less relevant in evaluating the partner's ability. This assumption of partial observability appears reasonable, given that participants show a higher willingness to collaborate with those who contribute more to solving the puzzle, as demonstrated in Figure 5.

5 Results 1: The effect of receiving corrections on willingness to collaborate

5.1 Main results

Table 2 presents the results from the regression specified in equation 1. Columns 1-4 and 9-10 include all participants, while columns 5-6 report results for male participants only, and columns 7-8 for female participants.

In column 1, when no controls for between-participant variation are included, the coefficient for good correction is underestimated. Column 2 shows that without controlling for the partner's contribution, the coefficient for bad correction is both large and negative: the estimate is -0.508 and statistically significant at the 1% level. This suggests that participants are 50.8 percentage points less willing to collaborate with partners who made a bad correction (i.e., a correction that moved the puzzle away from the solution). Furthermore, the coefficient for bad correction is 0.271 more negative than that for good corrections. In column 3, controlling for partner contribution, the coefficient on the female partner dummy is close to zero, suggesting that participants do not have a preference for male or female partners in the absence of corrections. Along with Figure 5, this indicates that participants are indifferent to a partner's gender if the contributions are the same and no corrections occur.

Looking at column 3, the coefficient on any correction is large and negative: -0.198, statistically significant at the 1% level. This implies that participants are 19.8 percentage points less willing

Table 2: The effect of receiving corrections on willingness to collaborate

Dependent variable:	Willing to collaborate (yes=1, no=0)									
Sample:	All				Male		Female		All	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Good correction	-0.208***	-0.238***		-0.204***		-0.168***		-0.229***		-0.168***
	(0.028)	(0.030)		(0.024)		(0.036)		(0.033)		(0.036)
Bad correction	-0.518***	-0.508***		-0.100***		-0.011		-0.172***		-0.011
	(0.031)	(0.034)		(0.036)		(0.052)		(0.047)		(0.052)
Any correction			-0.198***		-0.152***		-0.237***		-0.152***	
			(0.022)		(0.031)		(0.030)		(0.031)	
Female partner	-0.003	-0.001	0.008	0.009	0.016	0.016	0.002	0.004	0.016	0.016
	(0.016)	(0.017)	(0.014)	(0.014)	(0.021)	(0.021)	(0.018)	(0.018)	(0.021)	(0.021)
Partner's contribution			0.083***	0.084***	0.077***	0.080***	0.090***	0.089***	0.077***	0.080***
			(0.003)	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)	(0.003)	(0.004)
Good correction x Female										-0.062
										(0.048)
Bad correction x Female										-0.161**
										(0.070)
Any correction x Female									-0.085*	
									(0.044)	
Female partner x Female									-0.014	-0.012
									(0.028)	(0.028)
Partner's contribution x Female									0.012**	0.009
									(0.005)	(0.005)
Individual FE		✓	✓	✓	✓	✓	✓	✓	✓	✓
Good correction	0.310***	0.271***		-0.104**		-0.157**		-0.057		
−Bad correction	(0.048)	(0.052)		(0.045)		(0.065)		(0.061)		
Baseline mean	0.781	0.781	0.781	0.781	0.779	0.779	0.784	0.784	0.781	0.781
Baseline SD	0.414	0.414	0.414	0.414	0.416	0.416	0.412	0.412	0.414	0.414
Adj. R-squared	0.104	0.100	0.334	0.335	0.306	0.306	0.365	0.369	0.337	0.338
No. observations	3180	3180	3180	3180	1510	1510	1670	1670	3180	3180
No. individuals	464	464	464	464	220	220	244	244	464	464
No. corrections	495	495	495	495	244	244	252	252	495	495
No. good corrections	385	385	385	385	194	194	202	202	385	385
No. bad corrections	170	170	170	170	84	84	88	88	170	170

Notes: This table presents the regression results of equation 1. Columns 1-4 and 9-10 include all participants, columns 5-6 include male participants only, and columns 7-8 include female participants only. Baseline mean and standard deviation are participants' willingness to collaborate with male partners who do not make any corrections. Standard errors in parentheses are clustered at the individual level. Significance levels: * 10%, ** 5%, and *** 1%.

to collaborate with someone who corrected them, regardless of whether the correction benefited the game. To offset this negative effect, the partner's contribution would need to increase by 0.79 standard deviations.[22] For male participants (column 5), the corresponding coefficient is -0.152 (statistically significant at 1%), and for female participants (column 7), it is -0.237 (statistically significant at 1%). Women react marginally more negatively to corrections than men (column 9, p-value < 0.10). Overall, participants are less willing to collaborate with someone who corrected their move.

This negative response is problematic if participants do not differentiate between good and bad corrections. Column 4 shows that even good corrections lead to a negative response: the coefficient for good correction is -0.204, statistically significant at the 1% level. This suggests that participants are less willing to collaborate with those who corrected them, even when the correction was beneficial. For men, the corresponding coefficient is -0.168 (statistically significant at 1%) in

[22] This 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 an arithmetic average from panel B of Table 1, calculated as $(2.73 + 2.87)/2$.

column 6, and for women, it is -0.229 (statistically significant at 1%). The difference between men and women is statistically insignificant (column 10).

Moreover, column 4 shows that bad corrections also result in a negative response: the coefficient is -0.100 (statistically significant at the 1% level), but participants respond less negatively to bad corrections than to good ones by 0.104 percentage points (statistically significant at the 5% level). For men, the coefficient for bad corrections is statistically insignificant (column 6), while for women it is -0.172 (statistically significant at 1%), indicating that women respond more negatively to bad corrections (column 10). Thus, the more muted negative response to bad corrections compared to good ones is driven by male participants. However, as shown later in Figure 8, the gender difference in the response to bad corrections disappears once we restrict the sample to puzzles solved in 60 seconds or less and to puzzles where partner's contribution is between 3 to 5. This suggests that female participants may be more sensitive to time pressure (Shurchkov 2012), failing to distinguish between good and bad corrections when under time constraints. However, it is also possible that outliers in partner contributions drive these results. Regardless, this is not a robust finding.

These results suggest that participants are generally unwilling to collaborate with partners who corrected them, even when the correction was beneficial, and that this behavior is irrational. As shown later in Figure 8, this negative response to corrections does not fade over time, as restricting the analysis to earlier rounds (rounds 1-5) does not change the estimates meaningfully.

5.2 Mechanisms

One potential mechanism is feedback aversion: a good correction may serve as negative feedback about the recipient's ability, while a bad correction does not. A key prediction of the information avoidance literature is that individuals with a high self-image are more averse to negative feedback (Kőszegi 2006) and less likely to trust it (Eil and Rao 2011). I examine this hypothesis in Table 3, which presents the results of equation 1 with an interaction term for high-ability participants. High-ability participants are those who solved more than the median number of puzzles (8 or more) in Part 1 of the experiment, where the maximum number of puzzles solved was 15. Columns 1-2 and 7-8 include all participants, while columns 3-4 report results for male participants only and columns 5-6 for female participants.

High-ability participants are expected to better distinguish between good and bad corrections and thus respond less negatively to both. However, they are also more likely to be confident in their ability, making them more sensitive to negative feedback. In column 1, the coefficient on the interaction between any correction and the high-ability dummy is negative and statistically significant at the 5% level, indicating that high-ability participants are less willing to collaborate with someone who corrected them than low-ability participants. However, this negative response is driven by good corrections, as shown in column 2: the interaction term for bad corrections is close to zero, while that for good corrections is negative and statistically significant. These results suggest that high-ability participants are less willing to collaborate with someone who corrected their mistakes, but they do not react similarly to corrections of right moves. This pattern is

Table 3: The effect of receiving corrections on high- vs. low-ability participants

Dependent variable:	Willing to collaborate (yes=1, no=0)							
Sample:	All		Male		Female		All	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Good correction		-0.155***		-0.107***		-0.207***		-0.107***
		(0.030)		(0.041)		(0.042)		(0.041)
Bad correction		-0.100**		0.002		-0.202***		0.002
		(0.047)		(0.063)		(0.064)		(0.063)
Any correction	-0.153***		-0.097***		-0.213***		-0.097***	
	(0.028)		(0.037)		(0.041)		(0.037)	
Female partner	0.016	0.017	0.037	0.036	-0.006	-0.007	0.037	0.036
	(0.018)	(0.018)	(0.027)	(0.028)	(0.024)	(0.023)	(0.027)	(0.028)
Partner's contribution	0.084***	0.084***	0.079***	0.081***	0.090***	0.089***	0.079***	0.081***
	(0.003)	(0.003)	(0.004)	(0.004)	(0.005)	(0.005)	(0.004)	(0.004)
Good correction x High ability		-0.118**		-0.182**		-0.048		-0.182**
		(0.050)		(0.075)		(0.066)		(0.075)
Bad correction x High ability		0.001		-0.060		0.074		-0.060
		(0.072)		(0.109)		(0.095)		(0.109)
Any correction x High ability	-0.108**		-0.153**		-0.051		-0.153**	
	(0.044)		(0.064)		(0.061)		(0.064)	
Female partner x High ability	-0.015	-0.016	-0.046	-0.046	0.015	0.018	-0.046	-0.046
	(0.028)	(0.028)	(0.043)	(0.043)	(0.036)	(0.036)	(0.043)	(0.043)
Partner's contribution x High ability	-0.002	-0.001	-0.004	-0.003	-0.002	-0.001	-0.004	-0.003
	(0.005)	(0.006)	(0.007)	(0.008)	(0.007)	(0.007)	(0.007)	(0.008)
Good correction x Female								-0.100*
								(0.059)
Bad correction x Female								-0.205**
								(0.090)
Any correction x Female							-0.116**	
							(0.055)	
Female partner x Female							-0.043	-0.043
							(0.036)	(0.036)
Partner's contribution x Female							0.011*	0.008
							(0.006)	(0.007)
Good correction x High ability x Female								0.134
								(0.100)
Bad correction x High ability x Female								0.134
								(0.145)
Any correction x High ability x Female							0.103	
							(0.088)	
Female partner x High ability x Female							0.061	0.064
							(0.056)	(0.056)
Partner's contribution x High ability x Female							0.002	0.002
							(0.010)	(0.011)
Individual FE	✓	✓	✓	✓	✓	✓	✓	✓
Good correction x High ability		-0.119		-0.122		-0.121		
−Bad correction x High ability		(0.092)		(0.124)		(0.138)		
Good correction x High ability		-0.273***		-0.289***		-0.256***		
+Good correction		(0.040)		(0.063)		(0.051)		
Bad correction x High ability		-0.099*		-0.058		-0.128*		
+Bad correction		(0.055)		(0.089)		(0.070)		
Baseline mean	0.781	0.781	0.779	0.779	0.784	0.784	0.781	0.781
Baseline SD	0.414	0.414	0.416	0.416	0.412	0.412	0.414	0.414
Adj. R-squared	0.335	0.336	0.308	0.308	0.364	0.368	0.337	0.339
No. observations	3180	3180	1510	1510	1670	1670	3180	3180
No. individuals	464	464	220	220	244	244	464	464

Notes: This table presents the regression results of equation 1 where I interact the regressors with a high-ability participant dummy. Columns 1-2 and 7-8 include all participants, columns 3-4 male participants only, and columns 5-6 female participants only. Baseline mean and standard deviation are participants' willingness to collaborate with male partners who do not make any corrections. Standard errors in parentheses are clustered at the individual level. Significance levels: * 10%, ** 5%, and *** 1%.

driven by male participants (column 4), not female participants (column 6), aligning with literature that suggests men tend to be more overconfident than women (Croson and Gneezy 2009). Online Appendix Figure A3 shows that these results are robust across different specifications.[23]

Another potential mechanism is that receiving a correction signals a partner's problematic personality as a colleague. The act of correcting someone could be perceived as unpleasant, making participants less inclined to work with such individuals. However, this explanation does not account for the asymmetric response to good versus bad corrections. It is also unlikely to observe such behaviors in such a short experiment.

6 Results 2: The effect of receiving corrections from women on willingness to collaborate

6.1 Main results

Table 4 presents the regression results of equation 1, where I interact the regressors with the female partner dummy to allow the effects of correction to differ by the gender of the corrector. Columns 1-2 and 7-8 include all participants, while columns 3-4 report results for male participants only, and columns 5-6 for female participants only.

In column 1, the coefficient on the interaction between the partner's contribution and the female partner is near zero and statistically insignificant, a pattern consistent across both male (column 3) and female participants (column 5). As shown earlier in Figure 5, these results suggest that participants – both men and women – do not over- or underestimate women's contributions when selecting a collaborator. The coefficient on the interaction between any correction and the female partner is also statistically insignificant, slightly negative for male participants (column 3) and slightly positive for female participants (column 5), but the difference is not statistically significant (column 7).

However, a different picture emerges when we look at good and bad corrections separately in column 2. The interaction between good correction and female partner remains statistically insignificant, but the interaction between bad correction and female partner is positive and statistically significant. Moreover, while the combined coefficient for good correction and its interaction with the female partner is negative and statistically significant at the 5% level, the combined coefficient for bad correction and its interaction with the female partner is not statistically significant. This indicates that participants are less willing to collaborate with a female partner who corrected their mistakes but not necessarily less willing when a female partner corrected their right moves. These patterns are driven primarily by male participants (column 4), with no significant effects observed for female participants (column 6). Male and female participants' responses to a female partner's good correction differ significantly at the 10% level.

[23] The literature shows that in the absence of feedback about their ability, participants with high and low abilities update their beliefs in a similar way, albeit both groups significantly underweight negative feedback (Castagnetti and Schmacker 2022; Möbius et al. 2022).

Table 4: The effect of receiving corrections from women on willingness to collaborate

Dependent variable:	Willing to collaborate (yes=1, no=0)							
Sample:	All		Male		Female		All	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Good correction		-0.187***		-0.104*		-0.248***		-0.104*
		(0.035)		(0.053)		(0.045)		(0.053)
Bad correction		-0.176***		-0.104		-0.218***		-0.104
		(0.051)		(0.076)		(0.064)		(0.076)
Any correction	-0.203***		-0.125***		-0.260***		-0.125***	
	(0.031)		(0.045)		(0.042)		(0.045)	
Female partner	0.013	0.001	0.026	0.003	-0.001	-0.002	0.026	0.003
	(0.022)	(0.022)	(0.029)	(0.030)	(0.032)	(0.032)	(0.029)	(0.030)
Partner's contribution	0.084***	0.083***	0.078***	0.077***	0.090***	0.089***	0.078***	0.077***
	(0.004)	(0.004)	(0.005)	(0.006)	(0.006)	(0.006)	(0.005)	(0.006)
Good correction x Female partner		-0.035		-0.119*		0.035		-0.119*
		(0.044)		(0.067)		(0.057)		(0.067)
Bad correction x Female partner		0.144**		0.168		0.090		0.168
		(0.070)		(0.102)		(0.093)		(0.102)
Any correction x Female partner	0.009		-0.051		0.047		-0.051	
	(0.041)		(0.059)		(0.056)		(0.059)	
Partner's contribution x Female partner	-0.002	0.002	-0.001	0.006	-0.001	-0.001	-0.001	0.006
	(0.005)	(0.005)	(0.007)	(0.007)	(0.008)	(0.008)	(0.007)	(0.007)
Good correction x Female								-0.144**
								(0.070)
Bad correction x Female								-0.115
								(0.100)
Any correction x Female							-0.135**	
							(0.062)	
Female partner x Female							-0.027	-0.005
							(0.043)	(0.044)
Partner's contribution x Female							0.013	0.012
							(0.008)	(0.008)
Good correction x Female partner x Female								0.154*
								(0.088)
Bad correction x Female partner x Female								-0.078
								(0.138)
Any correction x Female partner x Female							0.097	
							(0.081)	
Partner's contribution x Female partner x Female							-0.001	-0.007
							(0.010)	(0.011)
Individual FE	✓	✓	✓	✓	✓	✓	✓	✓
Good correction x Female partner		-0.179**		-0.287**		-0.055		
−Bad correction x Female partner		(0.088)		(0.133)		(0.116)		
Good correction x Female partner		-0.222***		-0.223***		-0.212***		
+Good correction		(0.031)		(0.045)		(0.042)		
Bad correction x Female partner		-0.032		0.064		-0.128*		
+Bad correction		(0.048)		(0.066)		(0.067)		
Baseline mean	0.781	0.781	0.779	0.779	0.784	0.784	0.781	0.781
Baseline SD	0.414	0.414	0.416	0.416	0.412	0.412	0.414	0.414
Adj. R-squared	0.333	0.336	0.305	0.307	0.365	0.369	0.336	0.339
No. observations	3180	3180	1510	1510	1670	1670	3180	3180
No. individuals	464	464	220	220	244	244	464	464

Notes: This table presents the regression results of equation 1, where the regressors are interacted with the female partner dummy. Columns 1-2 and 7-8 include all participants, columns 3-4 include male participants only, and columns 5-6 include female participants only. Baseline mean and standard deviation are participants' willingness to collaborate with male partners who do not make any corrections. Standard errors in parentheses are clustered at the individual level. Significance levels: * 10%, ** 5%, and *** 1%.

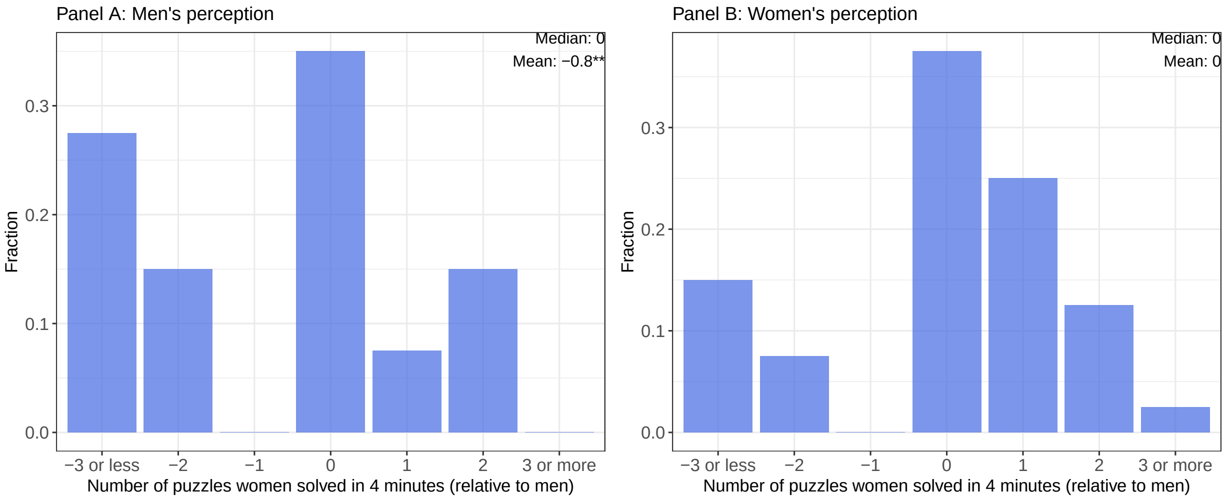
Thus, the results suggest that men may be less willing to collaborate with a woman who corrected their mistakes compared to a man, while there is no such asymmetry in response to corrections of right moves. For women, this asymmetric response to corrections by female versus male partners is not observed.

6.2 Mechanisms

Gender bias A possible explanation for these findings is gender bias: participants who are biased against women may react more negatively to female partners' good corrections. However, this mechanism does not seem to be driving the results. Online Appendix Table A2 presents the results of equation 1, where I interact the regressors with a high gender bias dummy. Participants with high gender bias are those whose gender bias, measured using six hostile and benevolent sexism questions from Karpowitz et al. (2024), is above the median for their gender. The interaction between good correction and the high gender bias dummy is statistically insignificant for all participants (column 2), male participants (column 4), and female participants (column 6). Similarly, the interaction between bad correction and the high gender bias dummy is also statistically insignificant. Thus, gender bias does not seem to explain the observed behavior.

Belief about women's puzzle-solving ability Another potential mechanism is participants' beliefs about women's puzzle-solving ability. While participants appear to equally prefer male and female collaborators with similar contributions in the absence of corrections (as shown in Figure 5 and Table 2), they may react differently to corrections due to stereotypes about women's abilities. For example, Sinclair and Kunda (2000) found that individuals are more likely to exhibit stereotypes against women when criticized by them, but not when praised. If participants believe that women are less competent at solving puzzles, corrections from women might damage participants' self-image more than corrections from men, leading to stronger negative reactions.

Figure 7: Belief about women's puzzle-solving ability



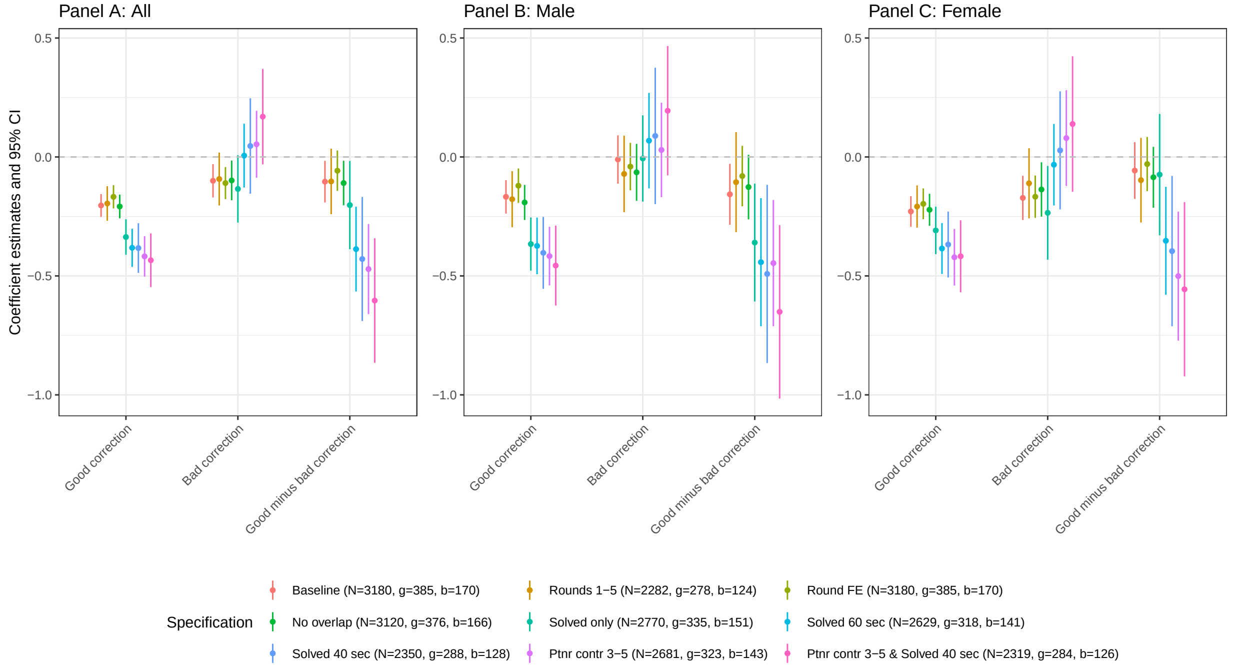
Notes: This figure shows men's (Panel A) and women's (Panel B) beliefs about women's puzzle-solving ability relative to men's, based on data from the follow-up experiment. Significance levels: * 10%, ** 5%, and *** 1%.

Figure 7 shows participants' beliefs about women's puzzle-solving ability relative to men's, based on data from the follow-up experiment. Panel A shows that men believe that men are slightly better at solving puzzles than women, with an estimated belief that women solved 0.8 fewer puzzles on average in Part 1 of the main experiment. In contrast, Panel B shows that women believe there is no significant gender difference in puzzle-solving ability. These beliefs align with the pattern that men, but not women, react more negatively to good corrections from female partners.

7 Robustness checks

7.1 Robustness checks for results 1

Figure 8: Robustness of the results in Table 2



Notes: This figure re-estimates and plots the main coefficient estimates (dots) and 95% confidence intervals (lines) from Table 2 under various specifications, with "Baseline" referring to the specifications in the table for comparison. Panel A shows estimates for column 4, Panel B for column 6, and Panel C for column 8. Sample sizes (N), good corrections (g), and bad corrections (b) for each specification are shown in parentheses.

Across-round imbalance Figure 6 showed that participants were less willing to collaborate, corrected their partners more frequently, and were less likely to solve puzzles in rounds 6 and 7, resulting in some imbalance in key variables across rounds. To address this, I re-estimated the coefficient estimates and 95% confidence intervals for good corrections, bad corrections, and good minus bad corrections from Table 2 without including data from rounds 6 or 7 (labeled "Rounds 1-5") and with round fixed effects ("Round FE") in Figure 8. Panel A shows that the estimates for all participants remain consistent with the baseline results in column 4, plotted in the figure as "Baseline," although the confidence intervals are slightly wider due to fewer observations. This pattern holds for both male (Panel B) and female participants (Panel C), confirming that the results are not driven by across-round imbalances.

Puzzles with both good and bad corrections In Section 3.1, I noted that 60 puzzles involved both good and bad corrections, creating potential ambiguity about which correction influenced participants' decisions. To test the robustness of the results to this concern, I re-estimated the coefficients without these overlapping puzzles (labeled "No overlap" in Figure 8). The estimates remain consistent with the baseline results, and this holds true for both male (Panel B) and female participants (Panel C), indicating that the results are not affected by these overlaps.

Unsolved puzzles In Section 3.1, I reported that 14% of puzzles were unsolved. To ensure these puzzles were not driving the results, I re-estimated the coefficients for good and bad corrections using only solved puzzles ("Solved only" in Figure 8). The coefficient for good corrections is more negative than in the baseline specification, but the overall pattern remains unchanged: participants are less willing to collaborate with those who corrected their mistakes, particularly male participants (Panel B) and to a lesser extent female participants (Panel C). Thus, unsolved puzzles do not drive the results.

Time to solve the puzzle It is possible that the time taken to solve the puzzle, rather than the correction itself, mediates the results. To test this, I re-estimated the coefficients for puzzles solved within 60 seconds ("Solved 60 sec") and 40 seconds ("Solved 40 sec") in Figure 8. The coefficient for good corrections remains more negative than in the baseline, while the coefficient for bad corrections approaches zero, particularly among female participants (Panel C). These findings suggest that time to solve the puzzle does not mediate the results and further confirm that participants do not respond rationally to corrections.

Outliers and functional form of partner contribution As shown in Figure 4, most partner contributions are between 3-5, but some outliers exist. Moreover, the empirical strategy assumes a linear relationship between partner contributions and perceived ability, which may not fully capture the effect. To address this, I re-estimated the coefficients for puzzles where partner contributions were between 3 and 5 ("Ptnr contr 3-5") and for puzzles both within this range and solved within 40 seconds ("Ptnr contr 3-5 & Solved 40 sec"). The estimates remain consistent with the baseline, and this pattern holds for both male (Panel B) and female participants (Panel C), indicating that outliers do not drive the results. These robustness checks suggest that being corrected, rather than the quality of the correction, drives the participants' unwillingness to collaborate, which is an irrational response.[24]

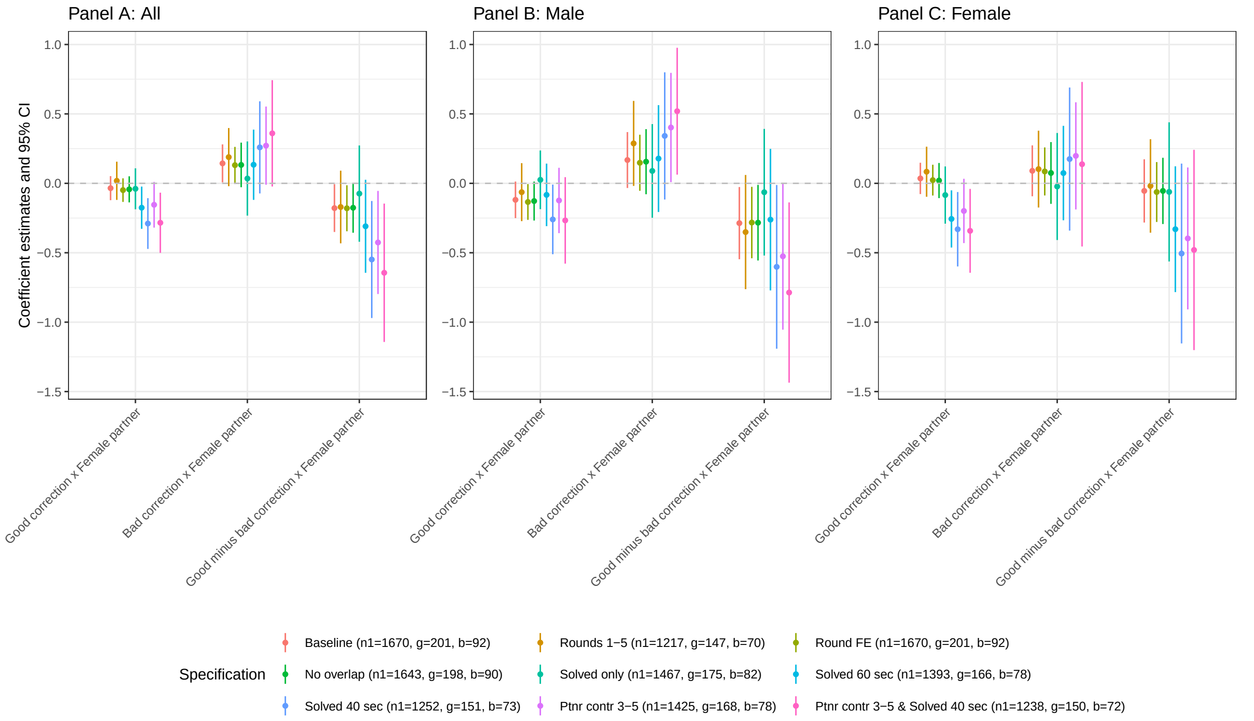
Finally, looking at the number of good and bad corrections in parenthesis next to each specification's name, there are enough of both types of corrections for each specification. So, the

[24] I did not present the specification where we restrict the sample to the puzzles with the partner contribution being exactly four because of the very wide confidence intervals that make other specifications difficult to read. However, the conclusions do not change: the coefficient estimates on good corrections are more negative than that of the baseline specification, close to the specification where we restrict the sample to solved puzzles only (the specification "Solved only"), albeit with wider confidence intervals. The coefficient estimate on bad correction is also more negative and statistically indistinguishable from the coefficient estimate on good correction.

(in)difference in the coefficient estimates on these specifications cannot be attributed to a lack of enough corrections.

7.2 Robustness checks for results 2

Figure 9: Robustness of the results in Table 4



Notes: This figure re-estimates and plots the main coefficient estimates (dots) and 95% confidence intervals (lines) of Table 4 under various specifications. The specification "Baseline" matches the table's specifications for comparison. Panel A plots the estimates for column 2, Panel B for column 4, and Panel C for column 6. Number of observations (n1), good corrections (g), and bad corrections (b) where the partner is female are indicated in parentheses.

Figure 9 presents the same robustness checks as Figure 8, confirming that the results in Table 4 are generally robust across various specifications. However, a few noteworthy differences emerge: (i) when the sample is restricted to puzzles solved within 40 seconds, the negative effect of female partners' good corrections becomes more pronounced for male participants; and (ii) when the sample is restricted to puzzles solved within 60 or 40 seconds, female participants also show a negative response to female partners' good corrections. The effect of female partners' bad corrections remains stable for female participants and becomes slightly more positive for male participants.

Overall, the asymmetric effect of female partners' good and bad corrections is primarily observed among male participants. For female participants, the asymmetric effect is not statistically significant across the specifications.

As in Figure 8, there are sufficient numbers of both types of corrections across samples, so the (in)significance of coefficients cannot be attributed to a lack of observations.

8 Conclusion

Teamwork is increasingly required in modern workplaces, but interpersonal frictions can inhibit its success. This paper highlights an important factor that undermines effective collaboration: individuals' reluctance to work with those who corrected them. I demonstrate that after receiving a correction, individuals are less willing to collaborate with the person who corrected them, even when it benefits the team.

The likely mechanism behind these behaviors is negative feedback aversion. Participants who are more confident in their ability are much less willing to collaborate with someone who corrected their mistakes, but not with those who confirmed their correct actions. Furthermore, I find suggestive evidence (at the 10% significance level) that men, but not women, are less willing to collaborate with women than with men who corrected their mistakes. Interestingly, the gender of the corrector does not influence collaboration decisions when corrected for right actions. These men's differential responses to women's corrections could be influenced by (inaccurate) beliefs about women's abilities, making collaboration in mixed-gender teams more challenging.

Of course, the effect sizes found in this study may vary in different real-world settings. Several factors absent from my experiment could amplify negative reactions to corrections in the workplace, including: (i) reputation costs (Bénabou and Tirole 2006), (ii) emotional stakes in the task, and (iii) hierarchical dynamics. First, the emotional cost of being corrected is likely to be higher when others witness it. Second, corrections in the workplace often relate to tasks individuals are deeply invested in, unlike the puzzle in my experiment. For instance, receiving critical feedback on a paper can feel far more personal than being corrected in a game-like setting. Third, since all participants in my experiment were equals, the dynamics of junior-senior relationships were not present. In a typical workplace, corrections from juniors to seniors might trigger stronger negative responses. Indeed, individuals tend to be more cautious when correcting senior colleagues.

On the other hand, factors that were not present in the experiment may also mitigate negative responses to corrections in real-world settings. These include (i) relationships and (ii) the ambiguity of corrections. First, participants in the study were strangers, whereas colleagues in workplaces typically know each other. A correction from someone with whom one has a positive relationship may be received more favorably – or more negatively if the relationship is strained. Second, in the experiment, corrections were clearly identifiable and binary, whereas in the workplace, corrections are often more subtle and nuanced. These considerations suggest that fostering strong relationships among colleagues may be crucial to minimizing negative reactions to corrections and promoting effective teamwork.

This paper provides a controlled experimental benchmark for studying these dynamics in collaborative environments. The findings underscore the importance of understanding the interpersonal effects of corrections, especially in mixed-gender teams, and highlight the potential for future research in more complex and hierarchical settings.

References

Abel, Martin. 2024. “Do Workers Discriminate against Female Bosses?” *Journal of Human Resources* 59 (2): 470–501.

Alan, Sule, Gozde Corekcioglu, and Matthias Sutter. 2023. “Improving Workplace Climate in Large Corporations: A Clustered Randomized Intervention.” *The Quarterly Journal of Economics* 138 (1): 151–203.

Bénabou, Roland, and Jean Tirole. 2006. “Incentives and Prosocial Behavior.” *American Economic Review* 96 (5): 1652–1678.

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.

Boskamp, Elsie. 2023. “35+ Compelling Workplace Collaboration Statistics: The Importance Of Teamwork.” *Zippia.*

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.

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

Chandrasekhar, Arun G., Benjamin Golub, and He Yang. 2019. *Signaling, Shame, and Silence in Social Learning.* Working Paper 3261632.

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

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

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

Cooper, Sarah. 2018. *How to Be Successful Without Hurting Men’s Feelings: Non-threatening Leadership Strategies for Women.* London, UK: Square Peg.

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

Eil, David, and Justin M. Rao. 2011. “The Good News-Bad News Effect: Asymmetric Processing of Objective Information about Yourself.” *American Economic Journal: Microeconomics* 3 (2): 114–138.

Erkal, Nisvan, Lata Gangadharan, and Boon Han Koh. 2023. “Do Women Receive Less Blame than Men? Attribution of Outcomes in a Prosocial Setting.” *Journal of Economic Behavior & Organization* 210:441–452.

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.

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

Guo, Joyce, and María P. Recalde. 2023. “Overriding in Teams: The Role of Beliefs, Social Image, and Gender.” *Management Science* 69 (4): 2239–2262.

Hardt, David, Lea Mayer, and Johannes Rincke. 2024. “Who Does the Talking Here? The Impact of Gender Composition on Team Interactions.” *Management Science.*

Heursen, Lea, Eva Ranehill, and Roberto A. Weber. 2022. *Are Women Less Effective Leaders than Men? Evidence from Experiments Using Coordination Games.* Working Paper.

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

ISTAT. 2024. “Contanomi - Quante bambine e quanti bambini si chiamano...? [Baby names - How many babies are named as...?]” Calcolatori. https://www.istat.it/dati/calcolatori/contanomi/.

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

Karpowitz, Christopher F., Stephen D. O’Connell, Jessica Preece, and Olga Stoddard. 2024. “Strength in Numbers? Gender Composition, Leadership, and Women’s Influence in Teams.” *Journal of Political Economy* 132 (9): 3077–3114.

Klinowski, David. 2023. “Voicing Disagreement in Science: Missing Women.” *The Review of Economics and Statistics,* 1–40.

Kőszegi, 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.

Möbius, Markus M., Muriel Niederle, Paul Niehaus, and Tanya S. Rosenblat. 2022. “Managing Self-Confidence: Theory and Experimental Evidence.” *Management Science* 68 (11): 7793–7817.

Shurchkov, Olga. 2012. “Under Pressure: Gender Differences in Output Quality and Quantity under Competition and Time Constraints.” *Journal of the European Economic Association* 10 (5): 1189–1213.

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

Thelwall, Mike, and Nabeil Maflahi. 2022. “Research Coauthorship 1900–2020: Continuous, Universal, and Ongoing Expansion.” *Quantitative Science Studies* 3 (2): 331–344.

Online Appendix

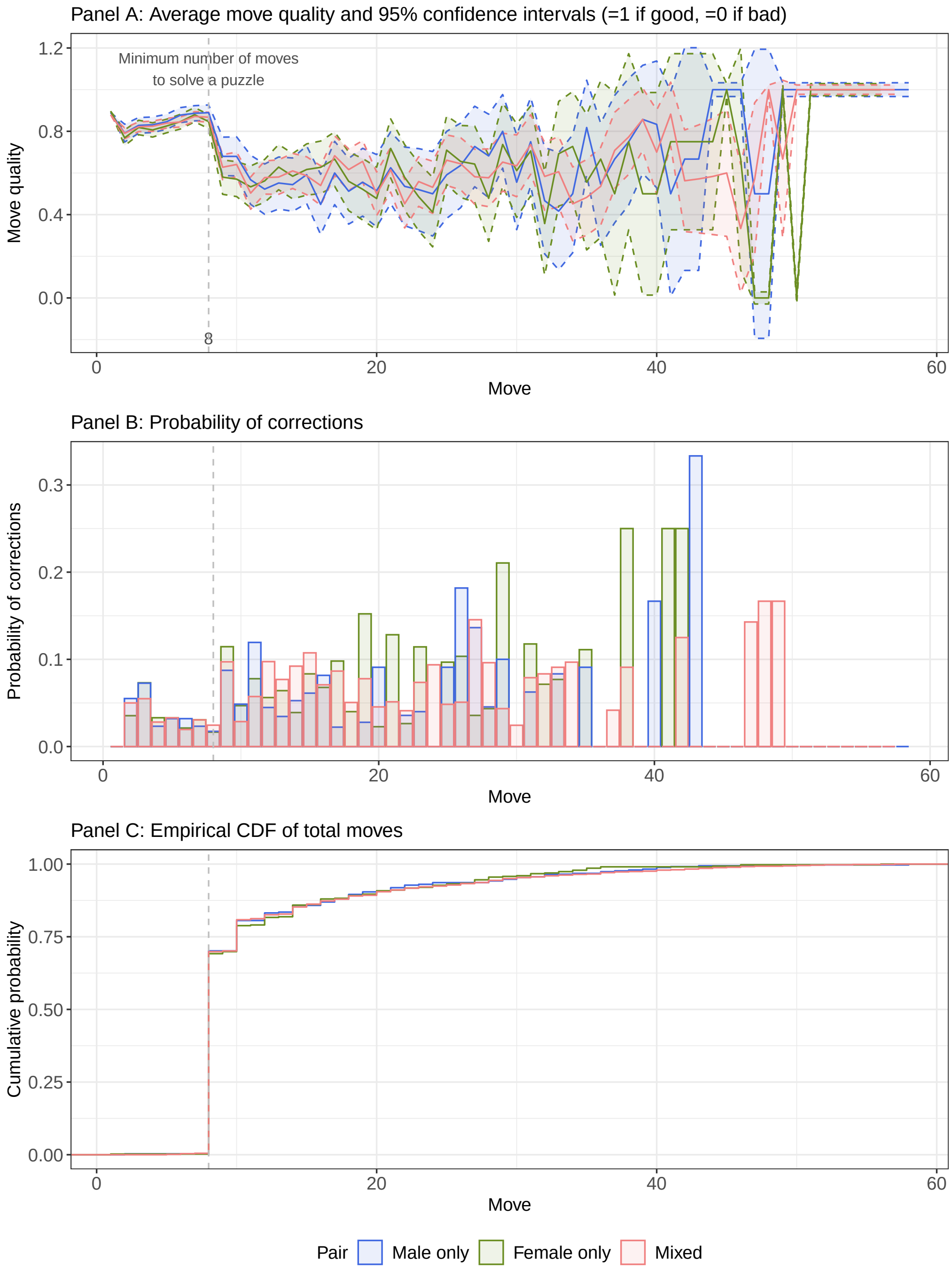
A Additional figures and tables

Table A1: Participants' characteristics

	Male (N=220)			Female (N=244)			Difference (Male – Female)	
	Mean	SD	Median	Mean	SD	Median	Mean	P-value
Age	25.87	4.33	25	24.45	3.13	24	1.41	0.00
Gender bias [0-1]	0.29	0.19	0.29	0.17	0.16	0.12	0.12	0.00
Region of origin (within Italy)								
North	0.36			0.32			0.04	0.37
Center	0.24			0.23			0.01	0.77
South	0.40			0.45			-0.06	0.23
Major:								
Humanities	0.22			0.45			-0.23	0.00
Social sciences	0.27			0.24			0.03	0.52
Natural sciences	0.20			0.12			0.08	0.02
Engineering	0.23			0.05			0.17	0.00
Medicine	0.08			0.13			-0.05	0.08
Program:								
Bachelor	0.26			0.34			-0.08	0.06
Master	0.68			0.63			0.05	0.26
Doctor	0.06			0.03			0.03	0.11

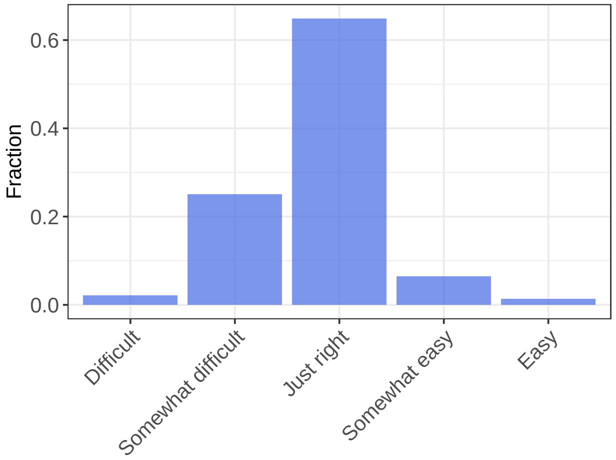
Notes: This table describes participants' characteristics. P-values of the difference between male and female participants are calculated with heteroskedasticity-robust standard errors.

Figure A1: Move quality, probability of corrections, and empirical CDF of total moves



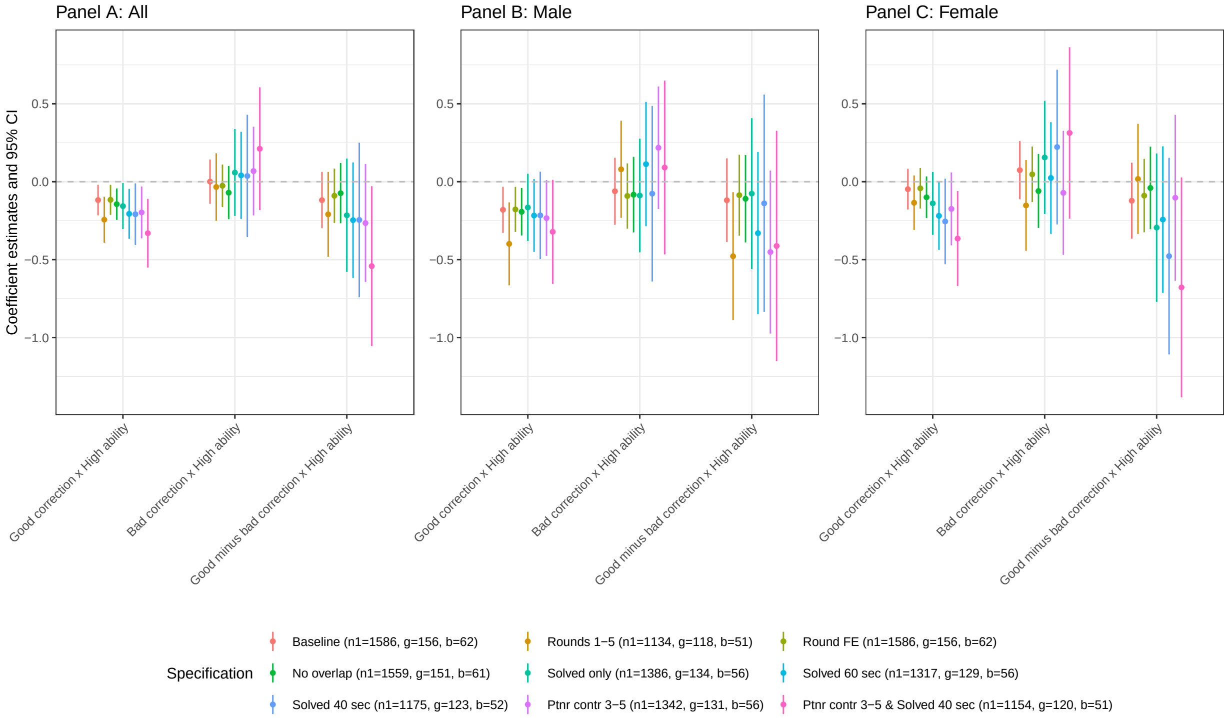
Notes: The average move quality along with 95% confidence intervals (panel A), the probability of corrections in each move (panel B), and the empirical CDF of total moves (panel C) separately for males only (blue), females only (green), and mixed gender pairs (red). The confidence interval of panel A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

Figure A2: Perceived puzzle difficulty



Notes: This figure shows the participants' perceived puzzle difficulty from the post-questionnaire.

Figure A3: Robustness of the results in Table 3



Notes: This figure re-estimates and plots the main coefficient estimates (dots) and 95% confidence intervals (lines) of Table 3 with various specifications, with the specification "Baseline" being the same as the specifications in the table for comparison. Panel A plots the estimates for column 2, Panel B for column 4, and Panel C for column 6. The number of observations where the participant is high ability (n1), the number of good corrections in n1 (g), and the number of bad corrections in n1 (b) in each sample are indicated in parenthesis next to the specification name and are based on all participants.

Table A2: Response to corrections made by women vs. men: Heterogeneity by gender bias

Dependent variable:	Willing to collaborate (yes=1, no=0)					
Sample:	All		Male		Female	
	(1)	(2)	(3)	(4)	(5)	(6)
Good correction		-0.180***		-0.066		-0.260***
		(0.046)		(0.071)		(0.058)
Bad correction		-0.196***		-0.201*		-0.185**
		(0.071)		(0.104)		(0.094)
Any correction	-0.215***		-0.137**		-0.267***	
	(0.041)		(0.065)		(0.053)	
Female partner	0.023	0.004	0.025	-0.006	0.022	0.014
	(0.029)	(0.029)	(0.041)	(0.042)	(0.039)	(0.040)
Partner's contribution	0.088***	0.087***	0.082***	0.079***	0.094***	0.094***
	(0.006)	(0.006)	(0.007)	(0.008)	(0.008)	(0.008)
Good correction x Female partner		-0.030		-0.151*		0.053
		(0.060)		(0.088)		(0.080)
Bad correction x Female partner		0.241**		0.312**		0.179
		(0.096)		(0.143)		(0.127)
Any correction x Female partner	0.036		-0.044		0.086	
	(0.055)		(0.079)		(0.076)	
Partner's contribution x Female partner	-0.005	0.001	-0.002	0.007	-0.007	-0.004
	(0.007)	(0.008)	(0.010)	(0.011)	(0.010)	(0.010)
Good correction x High gender bias		-0.018		-0.076		0.027
		(0.071)		(0.106)		(0.093)
Bad correction x High gender bias		0.036		0.173		-0.069
		(0.100)		(0.146)		(0.129)
Any correction x High gender bias	0.019		0.017		0.011	
	(0.063)		(0.091)		(0.089)	
Female partner x High gender bias	-0.027	-0.015	-0.010	0.008	-0.050	-0.039
	(0.044)	(0.044)	(0.058)	(0.059)	(0.064)	(0.065)
Partner's contribution x High gender bias	-0.009	-0.008	-0.010	-0.006	-0.008	-0.010
	(0.008)	(0.009)	(0.010)	(0.011)	(0.012)	(0.012)
Good correction x Female partner x High gender bias		-0.003		0.060		-0.025
		(0.089)		(0.134)		(0.118)
Bad correction x Female partner x High gender bias		-0.177		-0.257		-0.170
		(0.137)		(0.198)		(0.186)
Any correction x Female partner x High gender bias	-0.048		-0.011		-0.073	
	(0.082)		(0.118)		(0.114)	
Partner's contribution x Female partner x High gender bias	0.007	0.003	0.005	-0.000	0.011	0.008
	(0.011)	(0.011)	(0.014)	(0.014)	(0.016)	(0.016)
Individual FE	✓	✓	✓	✓	✓	✓
Baseline mean	0.782	0.782	0.780	0.780	0.784	0.784
Baseline SD	0.413	0.413	0.415	0.415	0.412	0.412
Adj. R-squared	0.333	0.335	0.304	0.305	0.363	0.368
No. observations	3173	3173	1503	1503	1670	1670
No. individuals	463	463	219	219	244	244

Notes: This table presents the regression results of equation 1 where I interact the regressors with a dummy for high gender bias participants. Columns 1-2 include all participants, columns 3-4 male participants only, and columns 5-6 female participants only. Baseline mean and standard deviation are participants' willingness to collaborate with male partners who do not make any corrections. Standard errors in parentheses are clustered at the individual level. Significance levels: * 10%, ** 5%, and *** 1%.

B Experimental instructions

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

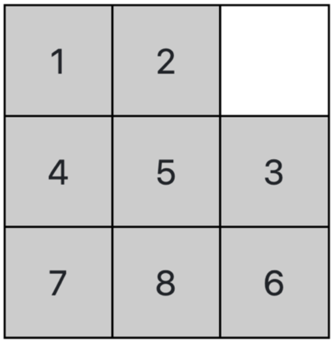
[I was chosen to participate] [I was chosen to leave]

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.



1	2	
4	5	3
7	8	6

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.

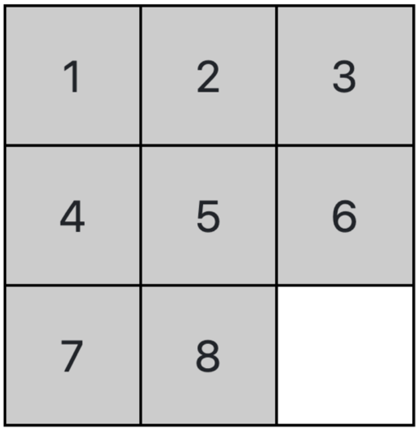
App: pt1

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:



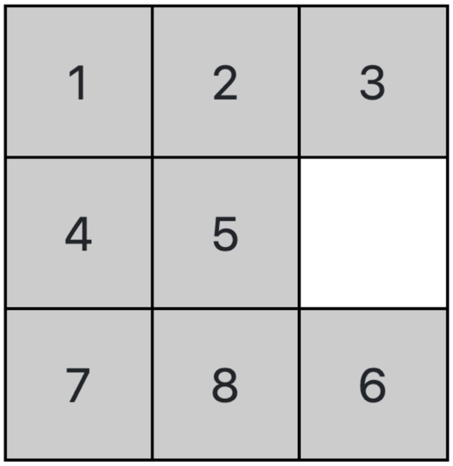
1	2	3
4	5	6
7	8	

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.



1	2	3
4	5	
7	8	6

puzzle_1.png

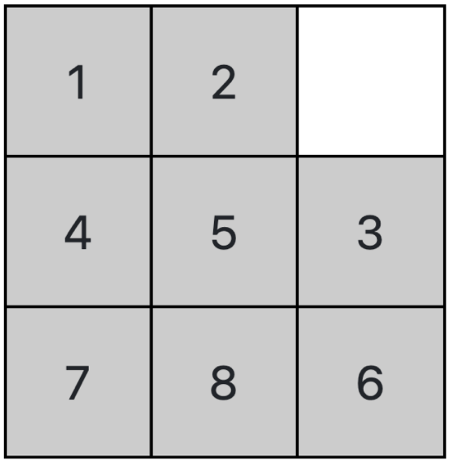
You can only move the tiles next to an empty cell and the tile you choose is moved to the empty cell. So, in this puzzle, there are 3 moves you can make: move 3 down, move 5 right, and move 6 up.

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.



1	2	
4	5	3
7	8	6

puzzle_2_0.png

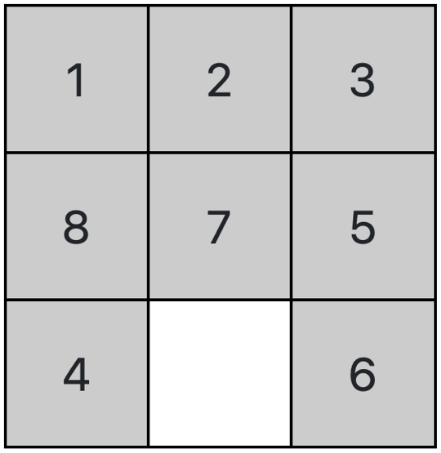
First, there are 2 moves you can make: move 2 right and move 3 up. Which moves should you make?

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.

<u>Example 3</u>:

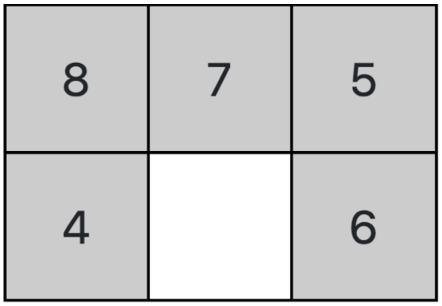
Finally, consider the following puzzle.



1	2	3
8	7	5
4		6

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.

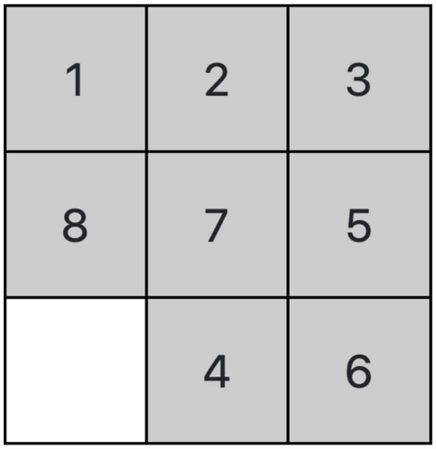


8	7	5
4		6

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?

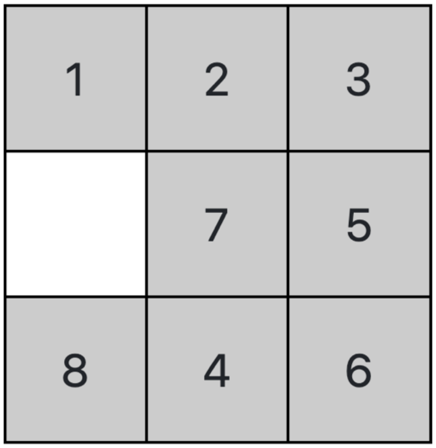
Let's try moving 4 right.



1	2	3
8	7	5
	4	6

puzzle_3_1_bad_0.png

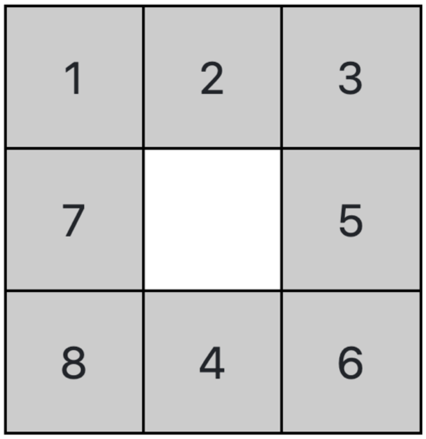
Now the only tile you can move is 8. So, let's move it down.



1	2	3
	7	5
8	4	6

puzzle_3_1_bad_1.png

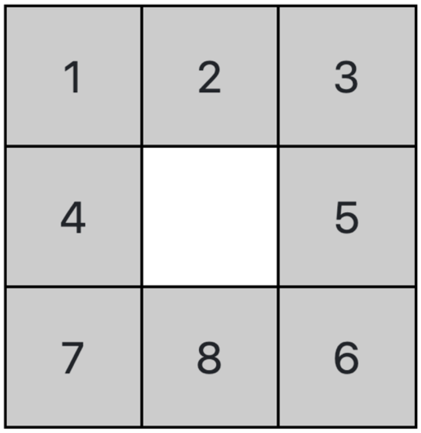
Now, if you ignore the top row which is already in the correct order, the only tile you can move is 7. So, let's move it to the left.



1	2	3
7		5
8	4	6

puzzle_3_1_bad_2.png

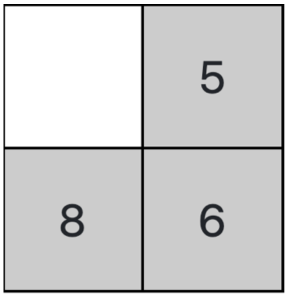
Then move 4 up, move 8 right, and move 7 down. Then you have made the left column in the correct order. You have moved tiles seven times until now.



1	2	3
4		5
7	8	6

puzzle_3_1_bad_3.png

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

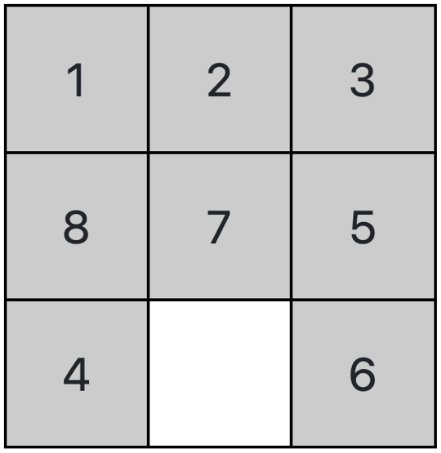


	5
8	6

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**.

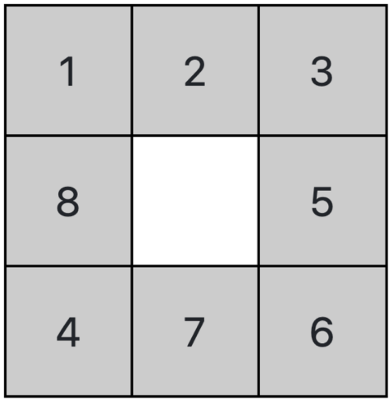
Let's go back to the initial puzzle.



1	2	3
8	7	5
4		6

puzzle_3_0.png

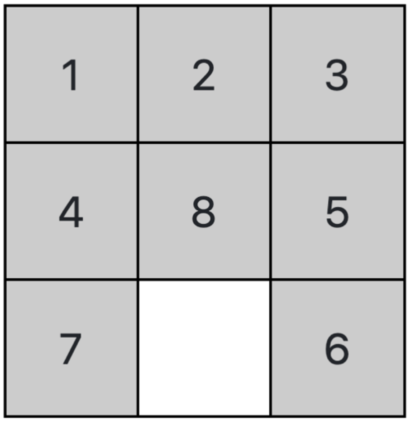
This time, let's try moving 7 down.



1	2	3
8		5
4	7	6

puzzle_3_1_good.png

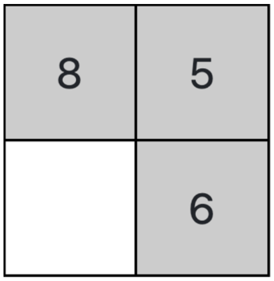
Then move 8 right, 4 up, and 7 left. Now you have made the left column in the correct order only with four moves.



1	2	3
4	8	5
7		6

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.

Understanding questions:

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

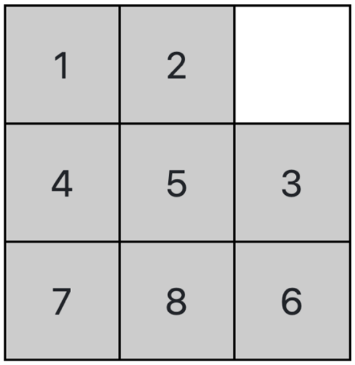
1. Which of the following statements is true?

- ✓In this part, 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

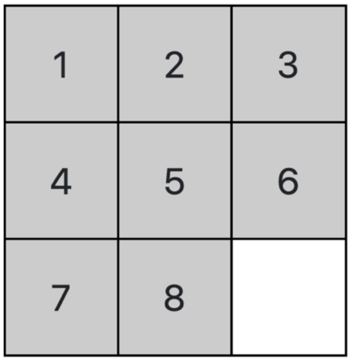
A



1	2	
4	5	3
7	8	6

puzzle_2_0.png

B



1	2	3
4	5	6
7	8	

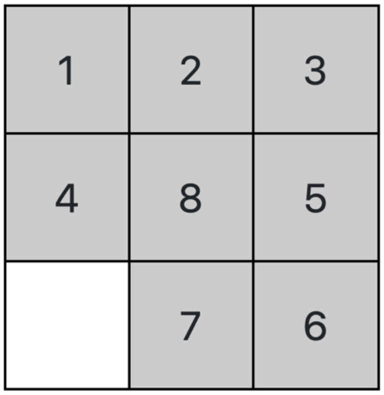
puzzle_goal.png

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.



1	2	3
4	8	5
	7	6

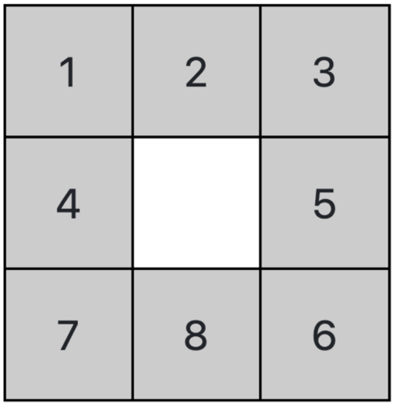
puzzle_3_3_good.png

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.



1	2	3
4		5
7	8	6

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.

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, 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.

App: pt4

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

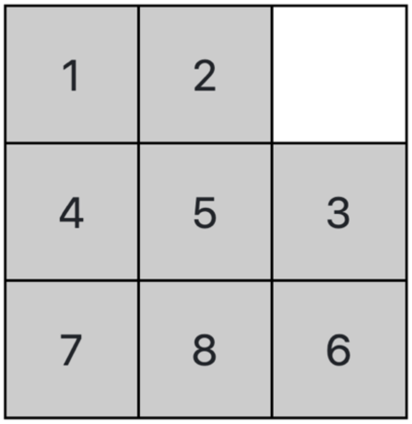
Thank you for showing up

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.

Welcome!

Thank you for participating in this study, which should take around 10 minutes.

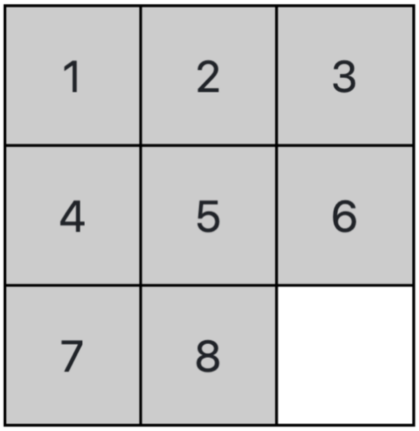
In this study, you will solve one puzzle, like the one shown below. After you solve the puzzle, we will ask you to guess about the puzzle. If your guess is correct, you will receive a bonus payment of £1.



1	2	
4	5	3
7	8	6

puzzle_2_0.png

To solve the puzzle, please move the tiles and order them as follows:



1	2	3
4	5	6
7	8	

puzzle_goal.png

To move a tile, please click it. It will move to the empty cell. You can only move the tiles adjacent to the empty cell.

<u>**Comprehension questions**</u>:

Before you proceed, please answer the following comprehension questions. Please re-read the instructions above if you are unsure about how to answer. After you answer, please click Next. **You have two opportunities to get these questions correct. If you cannot answer them in two attempts, you will be asked to return the survey and click "Stop Without Completing" on Prolific.**

1. How to move the tiles?

- Click the tile I want to move
- Drag the tile I want to move

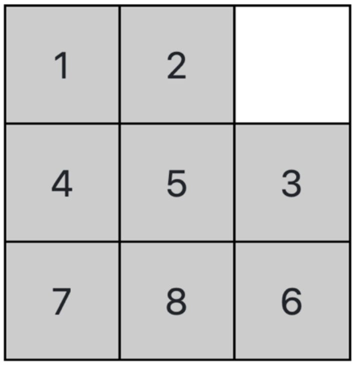
2. Which tiles can you move?

- Tiles adjacent to the empty cells
- Tiles on the top right

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

- A
- B

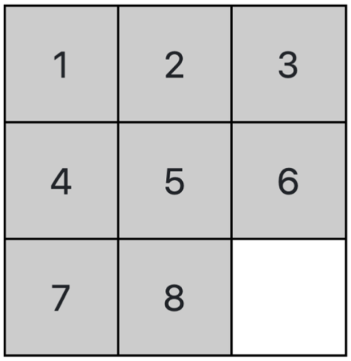
A



1	2	
4	5	3
7	8	6

puzzle_2_0.png

B



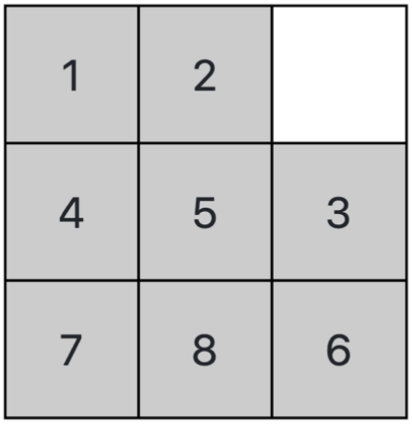
1	2	3
4	5	6
7	8	

puzzle_goal.png

[Next]

Puzzle

Solve the puzzle!



1	2	
4	5	3
7	8	6

Guess

Approximately 460 students at a university in Italy also solved the same puzzle you just solved but with different initial tile positions. They solved as many puzzles as possible within

4 minutes. On average, they solved 9 puzzles, with a minimum of 0 and a maximum of 15. The standard deviation is 2 puzzles. The decimals are rounded to the nearest integer.

Do you think there was a gender difference in the puzzle-solving ability among those 460 students? If so, which gender – male or female – solved more puzzles? If your guess is correct, you will receive a bonus of £1.

- Male students performed slightly better: they solved 1 more puzzle on average than female students.
- Male students performed better: they solved 2 more puzzles on average than female students.
- Male students performed significantly better: they solved 3 or more puzzles on average than female students.
- Female students performed slightly better: they solved 1 more puzzle on average than male students.
- Female students performed better: they solved 2 more puzzles on average than male students.
- Female students performed significantly better: they solved 3 or more puzzles on average than male students.
- Male and female students performed equally well: the difference is less than 1 puzzle on average.

[Next]

Thank you!

Thank you for your participation! Before you leave, can you tell us about yourself?

- Your gender: [Male, Female, Other]
- Your age: [Integer]
- 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]

[Next]

End of the study

The study is over. We will pay you £1.5 for your participation. If your guess is correct, we will pay you an additional £1 within 2 weeks.

Please click Next to complete the study.