Corrections and Gender in Team Collaboration

Yuki Takahashi[*]

May 26, 2025

Abstract

While successful teamwork often involves correcting colleagues' mistakes, it may have negative interpersonal consequences. In an experiment, I show that it also has negative economic consequences: individuals are less willing to collaborate with those who have corrected them, even when the correction benefits the team. The data are consistent with negative feedback aversion: individuals who initially received positive feedback about their ability are significantly less willing to collaborate with those who corrected their mistakes, but not with those who corrected their right actions. Additionally, I find that men, but not women, are more tolerant of women who corrected their right actions. It is potentially due to men's beliefs about women's abilities, making women's corrections of their right actions less ego-threatening. This reluctance to work with those who provide corrective feedback can undermine teamwork, and mixed-gender teams may attract less competent women due to gendered sorting.

JEL Classification: M54, D91, J16, C92

Keywords: Correction, Collaboration, Teamwork, Gender, 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. 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 A3) 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 (Jones 2021). However, as teamwork involves human interactions, several factors can inhibit its efficiency. 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 correcting colleagues' mistakes, which is essential for team success, but may also create interpersonal frictions. For example, a worker might need to correct a colleague's miscalculated numbers in a presentation slide or a flawed conclusion in a report. These corrections can damage relationships if taken personally. Further, 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). If these are the case, corrections may have economic consequences, which has not received much attention in the literature.

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 with female correctors. 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 an experiment 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 the puzzle progress). The task also provides a clear definition of a correction – reversing a partner's move – making it comparable across participants. At the 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 can differentiate good and bad moves: they are more likely to select as collaborators those who contributed more to solving the puzzle. I also find that men and women contribute equally, and that in the absence of corrections, participants are equally likely to select male and female collaborators with comparable contributions.

However, after controlling for contribution, participants are less willing to collaborate with those who corrected them, even when the corrections are beneficial to the team. This effect is substantial: it reduces the likelihood of collaboration by about 20 percentage points, or approximately 26% relative to the baseline mean, and would require an additional contribution of 0.84 standard deviations to offset.

The data are consistent with negative feedback aversion (Kőszegi 2006; Eil and Rao 2011): participants who receive positive feedback about their puzzle-solving ability are significantly less willing to collaborate with those who corrected their mistakes, but not with those who corrected their right moves.

Regarding the gender dynamics, I find that men respond differently to corrections depending on the gender of the corrector: while they respond less negatively to bad corrections from women, they react equally negatively to good corrections from both women and men. One possible explanation is that bad corrections by women are less threatening to men's self-image, as they are consistent with men's belief that women are less competent; by contrast, good corrections by women may challenge this belief. In line with this conjecture, men do in fact believe that women solved fewer puzzles than men in the individual practice round.

While this may sound contradictory to the fact that men do not prefer male over female collaborators in the absence of corrections, it is consistent with motivated stereotyping, which states that individuals hold stereotyping only when doing so protects their ego (Sinclair and Kunda 2000). Women, on the other hand, do not differentiate between male and female correctors and believe that women and men performed equally in the task.

Taken together, these findings suggest that the reluctance to collaborate with those who corrected them can indeed be another obstacle to efficient teamwork, and mixed-gender teams may attract less competent women due to gendered sorting.

The main contribution of this paper is to demonstrate that individuals' reluctance to collaborate with those who corrected them can hinder teamwork, an insight that adds to the literature on barriers to successful teamwork. Prior work shows that individuals are less willing to seek advice from senior colleagues due to concerns about signaling low ability (Chandrasekhar, Golub, and Yang 2019), and that they are less inclined to reciprocate colleagues 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), and even when they do, their ideas are less likely to be taken up by teams (Coffman, Flikkema, and Shurchkov 2021). Additionally, men tend to dominate team discussions, even when their abilities are lower than those of female teammates, which can reduce overall team performance (Hardt, Mayer, and Rincke

2024). Building on this literature, I show that gender dynamics influence how corrective feedback is received in team settings and how this can affect individuals' sorting into teams.

This paper also contributes to the literature on gender disparities in teamwork. Guo and Recalde (2023) find that individuals are more likely to override women's opinions than men's in a male-typed task. I extend Guo and Recalde by showing that men are more forgiving of women's mistakes but not women's right actions in a slightly male-typed task, leading to gendered sorting into teams. Relatedly, Isaksson (2018), using the same puzzle, finds that women claim less credit for their contributions than men, particularly in difficult puzzles, and that men are more likely to correct their partners. Corroborating these findings, Klinowski (2023) finds that female scientists are less likely than their male peers to criticize or correct other scientists' work.

Finally, this paper contributes to the literature on feedback aversion by providing suggestive evidence that individuals are averse to negative feedback – particularly those who initially received positive feedback – in a teamwork context. Existing research shows that individuals deviate from Bayesian updating in response to negative feedback about traits they value, such as IQ or attractiveness, while they do not deviate much in response to positive feedback (Eil and Rao 2011). In educational settings, positive feedback has been shown to reduce the grades of students who underestimated their academic performance, whereas negative feedback slightly improves the performance of those who initially overestimated themselves (Azmat et al. 2019).

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

The experiment was designed to collect data on collaborator preferences and other key variables and was conducted in a quasi-laboratory format.[1] Participants and experimenters were connected via Zoom throughout the session. Participants' cameras and microphones were turned off except at the beginning of the session, and participants completed the tasks remotely using their personal computers. Aside from the remote setup, the experiment followed the procedures of a traditional physical laboratory setting.

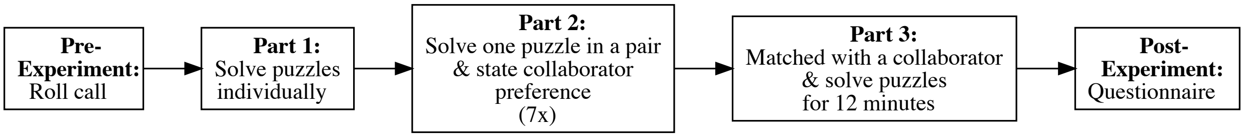
2.1 Experimental Design

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,

[1] The experimental instructions for both this experiment and the follow-up experiment are available in Appendix B.

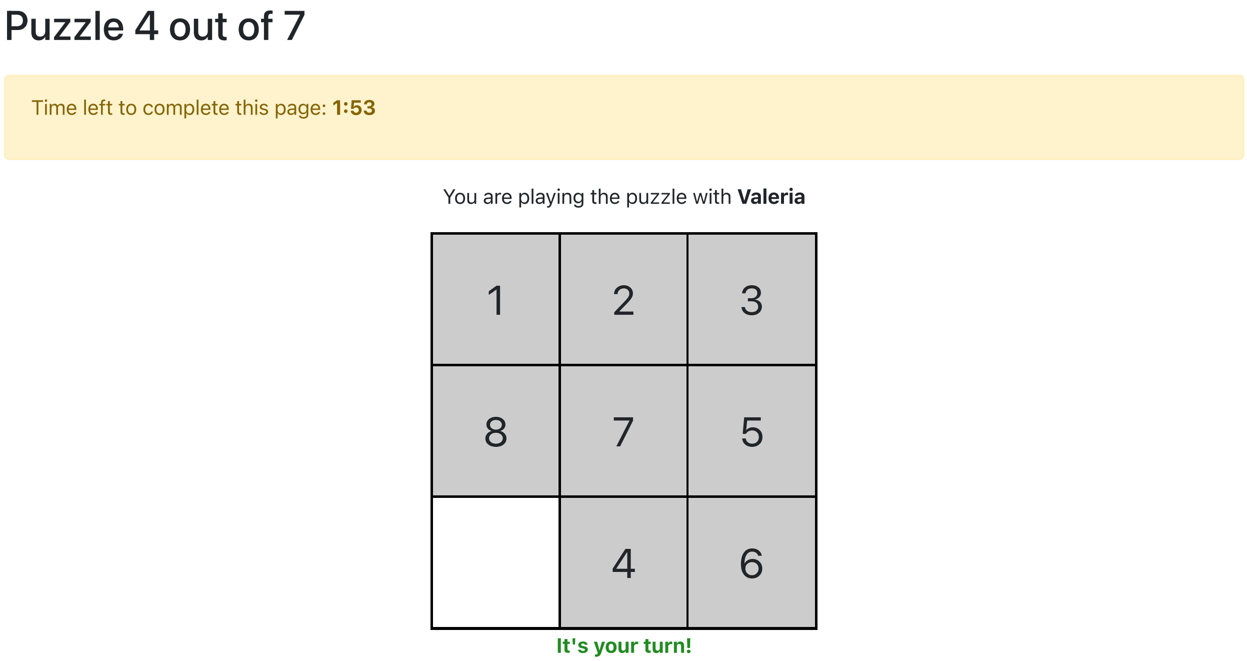
participants learned the rules of Part 3 and stated their collaborator preferences after solving one puzzle with each potential collaborator. In Part 3, participants worked on puzzles with collaborators selected based on their preferences from Part 2. At the beginning of each part, participants answered 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 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

The Team Task

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.

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 immediately afterward. This definition allowed for an objective comparison of corrections between participants, as corrections themselves counted as moves.[2] This setup captures 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

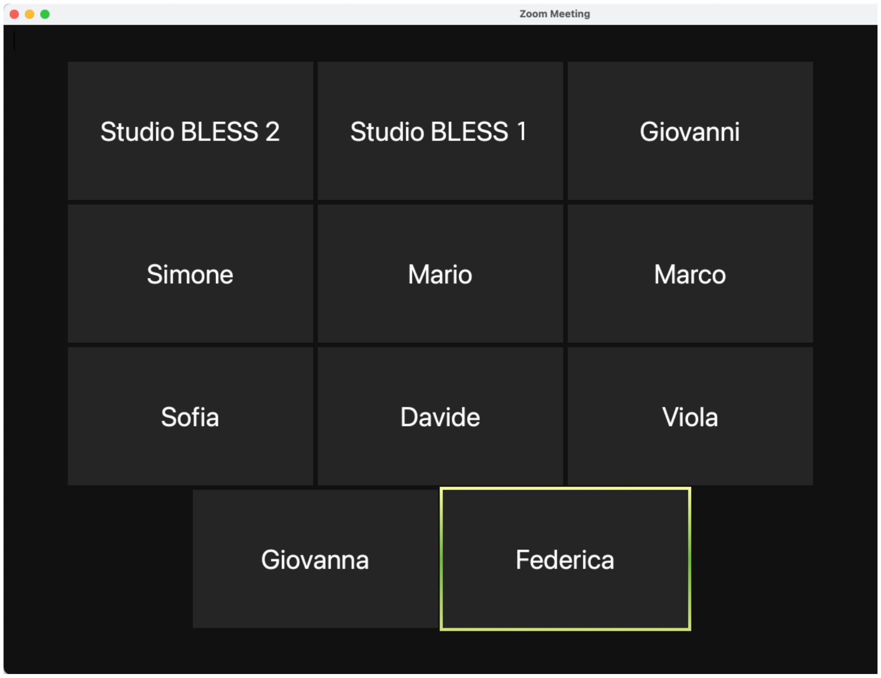
[2] Because some corrections happened early in the puzzle and others later, I capture the average effect of a correction in the analysis.

[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).

Figure 3: Zoom Screen



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

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.

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

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

[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.

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.

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.

2.2 Implementation and Participant Characteristics

The experiment was programmed using oTree (Chen, Schonger, and Wickens 2016) and conducted in Italian during 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 who (i) were born in Italy and (ii)

[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.

[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.

had not previously participated in gender-related experiments.[10,11] The first condition was 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] The second condition was imposed to reduce potential experimenter demand effects.

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

Appendix Table A1 summarizes participants' demographic and academic characteristics. On average, participants were 25 years old, and 96% were either bachelor's or master's students, with only a small number of PhD students. No economics PhD students participated. Nearly 60% of participants majored in the humanities or social sciences, and the remaining 40% in the natural sciences, engineering, or medicine. Some gender differences emerged: male participants were slightly older (by 1.41 years) and exhibited marginally higher gender bias scores (by 0.12 points) than female participants. Men were also more likely to major in natural sciences and engineering and less likely to major in the humanities – a pattern observed in most OECD countries (see, for example, Carrell, Page, and West 2010). In the analysis, I control for potential differences associated with participants' major by including individual fixed effects.

2.3 Follow-Up Experiment

To collect data on participants' perceptions of the genderness of the puzzle used in the main experiment, I conducted a follow-up experiment. This experiment consisted of 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 – they believed had solved more puzzles in Part 1 of the main experiment, using a 7-point Likert scale. Participants who guessed correctly earned an additional £1 on top of the completion fee. Afterward, they completed a short demographic questionnaire.

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 comparability with the original participant pool, participants were required to be students residing 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 £1.50 completion fee.

[10] Sixteen participants from a pilot session with slightly different instructions are included in the analysis. The results are robust to excluding these participants.

[11] The laboratory prohibits deception, and no participants took part in experiments involving 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.

3 Data

I use data from Part 2 of the experiment, where collaborator preferences were recorded. Move-level data for each puzzle was aggregated to link participants' puzzle behavior with their collaborator preferences.[13]

3.1 Data Description

Table 1: Own and Partner's Puzzle Behaviors and Puzzle Outcomes

	All (N=3190)		Male (N=1507)		Female (N=1683)		Difference (Male – Female)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SE
Panel A: Own behaviors								
Contribution	3.00	2.86	3.09	2.70	2.92	3.00	0.18*	0.10
# puzzles solved in part 1	8.56	2.36	8.77	2.32	8.37	2.39	0.44**	0.22
Any correction	0.16	0.37	0.16	0.37	0.16	0.37	0.00	0.01
Good correction	0.12	0.33	0.12	0.33	0.12	0.33	0.00	0.01
Bad correction	0.06	0.23	0.06	0.23	0.06	0.23	0.00	0.01
(Fraction of female partners)	0.53	0.50	0.55	0.50	0.51	0.50	0.04**	0.02
Panel B: Partner's behaviors								
Contribution	3.00	2.86	3.03	2.92	2.97	2.81	0.06	0.10
# puzzles solved in part 1	8.56	2.36	8.55	2.39	8.58	2.34	-0.02	0.16
Any correction	0.16	0.37	0.16	0.36	0.16	0.37	0.00	0.01
Good correction	0.12	0.33	0.12	0.33	0.13	0.33	-0.01	0.01
Bad correction	0.06	0.23	0.05	0.22	0.06	0.23	-0.01	0.01
Panel C: Puzzle outcomes								
Willing to collaborate (yes=1, no=0)	0.73	0.45	0.72	0.45	0.73	0.44	-0.01	0.02
Willing to collaborate (residualized)	0.00	0.41	0.00	0.41	0.00	0.42	0.00	0.00
Time spent (second)	42.09	34.82	41.41	34.31	42.70	35.26	-1.29	1.25
Total moves	11.41	7.91	11.48	8.22	11.35	7.62	0.13	0.30
Puzzle solved	0.87	0.34	0.88	0.33	0.86	0.34	0.01	0.01
Consecutive correction	0.04	0.20	0.05	0.21	0.04	0.20	0.00	0.01

Notes: This table summarizes participants' own puzzle behaviors (Panel A), their partners' behaviors (Panel B), and puzzle outcomes (Panel C). P-values for gender differences are calculated with standard errors clustered at the individual level. Contribution is defined as the net good moves in a given puzzle (i.e., the number of good moves minus the number of bad moves). Significance levels: * 10%, ** 5%, and *** 1%.

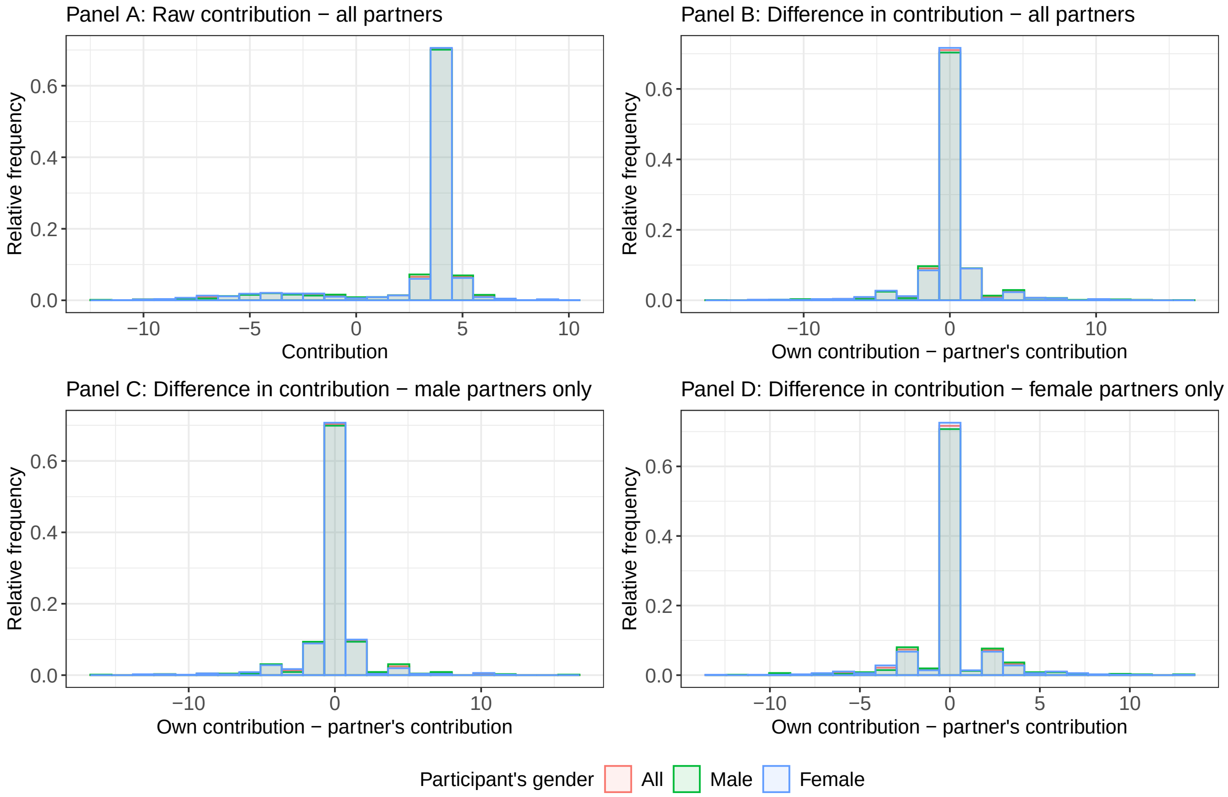
Table 1 summarizes participants' puzzle behaviors and outcomes. Panel A shows that, on average, participants solved 8.6 puzzles in Part 1, contributed 3 net good moves in Part 2, and corrected their partners in 16% of puzzles (512 out of 3,190). Of these corrections, 12% were good and 6% were bad, with both types of corrections occurring in about 2% of puzzles. Although there

[13] 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.

are gender differences in puzzle-solving ability, they are quantitatively small: men made 0.2 more good moves in Part 2 (p-value < 0.10) and solved 0.4 more puzzles in Part 1 (p-value < 0.05) than women.[14] These findings are consistent with Isaksson (2018), who reported no significant gender differences in contributions or puzzle-solving using the same task. Regarding correction behavior, men and women were similarly likely to correct their partners, which contrasts with the findings of Isaksson (2018), who observed that men correct more often, and Klinowski (2023), who found that men are more likely to point out and penalize mistakes.

Of the 512 puzzles in which at least one correction occurred, 367 (72%) involved only one correction, while 145 (28%) involved more than one. Among those 145 puzzles, 74 (51%) experienced only good corrections, 10 (7%) only bad corrections, and 61 (42%) included both. I later show that the results are robust to excluding puzzles with overlapping corrections. Finally, the last row of Panel A shows that male participants were slightly more likely to be paired with female partners, a difference of about three percentage points.

Figure 4: Distribution of Contributions



Notes: This figure presents the distribution of individual contributions for all participants and by gender. Panel A shows the raw contribution distribution, while Panels B–D depict the difference in contributions between participants and their partners, with Panel C focusing on male partners and Panel D on female partners. Contribution is defined as the net good moves made in a given puzzle (the number of good moves minus the number of bad moves an individual made).

[14] The correlation coefficient between contributions and the number of puzzles solved in Part 1 is 0.1043, with a p-value below 0.001 (standard errors clustered at the individual level).

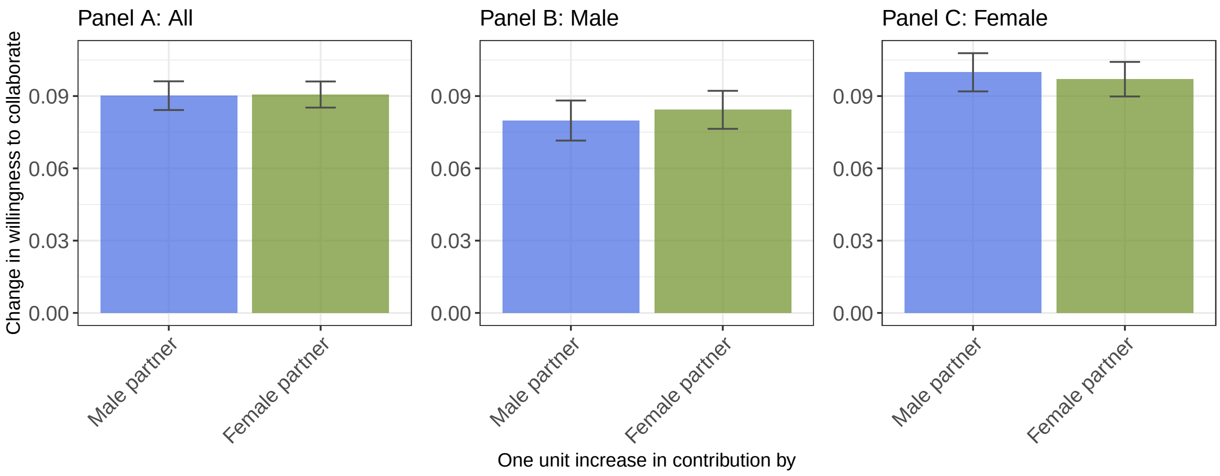
Building on Panel A of Table 1, Panel A of Figure 4 shows that most participants contributed similarly, regardless of gender: around 70% of puzzles resulted in 4 net good moves, and about 90% of puzzles fell within the 3–5 range. Panels B–D confirm that these patterns hold across different gender pairings. Although some outliers are present, I later show that excluding them does not alter the overall results.

Panel B of Table 1 shows that puzzle-solving ability and the likelihood of correcting a partner's move (whether correct or incorrect) were consistent across gender pairings, suggesting that random assignment was successful. Participants were corrected in 16% of puzzles, with 12% of those corrections classified as good and 6% as bad. These percentages do not sum to 16% because some puzzles included both types of correction. I later show that the results are robust to excluding such overlapping corrections.

Panel C shows that participants chose to collaborate with their partners in 71% of puzzles, and there is sufficient within-subject variation.[15] On average, participants spent 43 seconds per puzzle (out of a maximum of 120 seconds) and made 11 moves. In 87% of puzzles, participants successfully solved the puzzle, and in 4% they corrected their partner's moves consecutively. Notably, when consecutive corrections occurred, the likelihood of selecting the partner as a collaborator dropped from 75% to 27%. There were no gender differences in these outcomes, further indicating that gender did not drive any imbalances. I later show that the results are robust to excluding unsolved puzzles and puzzles where participants took 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 the change in willingness to collaborate for all participants (Panel A), men (Panel B), and women (Panel C) following a one-unit increase in the contribution of male (blue) and female (green) partners, along with 95% confidence intervals calculated using standard errors clustered at the individual level.

[15] This variation is measured by the standard deviation of residualized willingness to collaborate, obtained by regressing willingness on individual fixed effects and using the residuals.

Figure 5 shows that participants – both men and women – respond positively to their partner's contribution, regardless of the partner's gender. Specifically, both male and female participants are more likely to collaborate with partners who contribute more. I show later that participants do not exhibit a preference for male or female collaborators in absence of corrections. The smaller increase in men's willingness to collaborate may stem from overconfidence, leading them to underestimate their partner's ability. The overall positive response to contributions indicates that participants correctly understand the notion of good and bad moves.[16]

3.3 Across-Round Balance

Figure 6 shows that key variables are generally balanced across rounds. However, some imbalance emerges in rounds 6 and 7, where participants are less willing to collaborate, correct their partners more frequently, and are less likely to solve the puzzle. I later show that the results are robust to excluding these rounds.[17]

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, the number of good moves minus the number of bad moves).
- ϵ_{ij}: the error term.

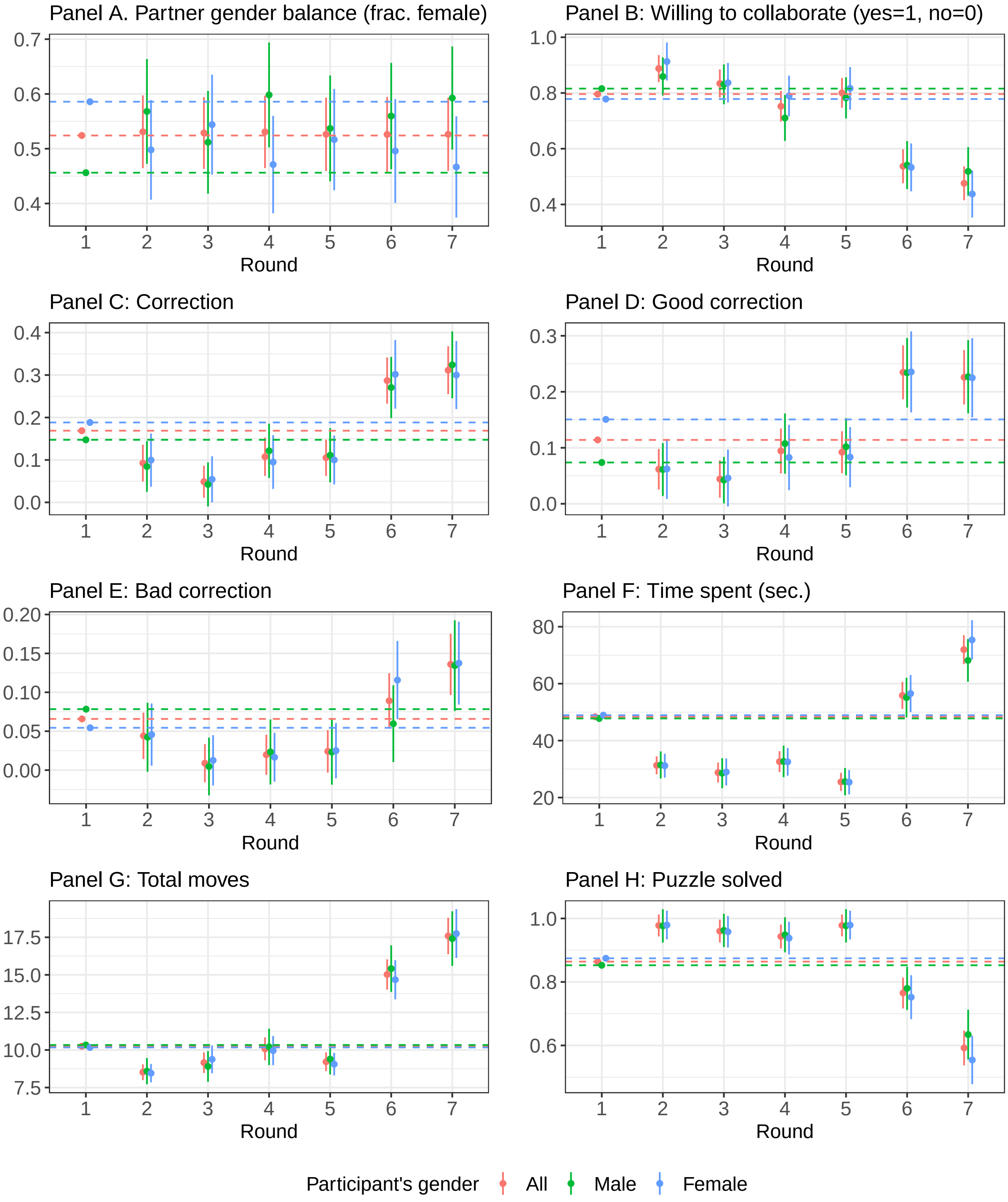
$\mu_i \equiv \sum_{k=1}^{N} \mu^k \mathbb{1}[i = k]$ represents the individual fixed effects, where N is the total number of participants, and $\mathbb{1}$ is the indicator variable. Standard errors are clustered at the individual level.[18]

[16] This is also supported by the high puzzle-solving rate and the tendency to favor high-contributing partners. See Appendix Figure A2 for participants' perceived puzzle difficulty from the post-experiment questionnaire, broken down for all participants (Panel A), male participants (Panel B), and female participants (Panel C).

[17] One exception is round 1, where male participants were more likely to have female partners. This imbalance, however, does not persist in rounds 2–7.

[18] 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 for all (red), male (blue), and female (green) participants: $y_{ij} = \beta_1 + \sum_{k=2}^{7} \beta_k \mathbb{1}[t_{ij} = k] + \epsilon_{ij}$, where $t_{ij} \in 1, 2, 3, 4, 5, 6, 7$ indicates the puzzle round played by participants i and j, $\mathbb{1}$ is an indicator function, and y_{ij} is the dependent variable indicated in each panel. I add β_1 to the estimates of β_2–β_7 for easier interpretation. Standard errors are clustered at the individual level.

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 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: Individuals Are Less Willing to Collaborate with Those Who Corrected Them

5.1 Main Results

First, looking at columns 1 and 2 of Appendix Table A2, the coefficient for bad corrections is large and negative when the partner's contribution is not controlled for: the estimate ranges from -0.525 to -0.513 and is statistically significant at the 1% level. This suggests that participants are 51.3 to 52.5 percentage points less willing to collaborate with partners who made a bad correction (that is, a correction that moved the puzzle away from the solution). Moreover, the coefficient for bad corrections is 0.267 to 0.303 more negative than that for good corrections. These patterns hold for both men (columns 3 and 4) and women (columns 5 and 6), with no statistically significant gender differences. This implies that participants can distinguish between good and bad corrections.

Table 2: 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)
Good correction		-0.209***		-0.181***		-0.230***		-0.181***
		(0.024)		(0.035)		(0.032)		(0.035)
Bad correction		-0.095***		-0.023		-0.153***		-0.023
		(0.035)		(0.051)		(0.046)		(0.051)
Any correction	-0.202***		-0.171***		-0.229***		-0.171***	
	(0.022)		(0.031)		(0.030)		(0.031)	
Female partner	0.004	0.005	0.009	0.009	0.001	0.003	0.009	0.009
	(0.014)	(0.014)	(0.021)	(0.021)	(0.018)	(0.018)	(0.021)	(0.021)
Partner's contribution	0.084***	0.085***	0.077***	0.079***	0.091***	0.090***	0.077***	0.079***
	(0.003)	(0.003)	(0.003)	(0.004)	(0.003)	(0.004)	(0.003)	(0.004)
Good correction x Female								-0.049
								(0.048)
Bad correction x Female								-0.130*
								(0.069)
Any correction x Female							-0.059	
							(0.043)	
Female partner x Female							-0.008	-0.007
							(0.028)	(0.028)
Partner's contribution x Female							0.014***	0.011**
							(0.005)	(0.005)
Individual FE	✓	✓	✓	✓	✓	✓	✓	✓
Good correction		-0.114***		-0.158**		-0.077		
−Bad correction		(0.044)		(0.064)		(0.059)		
Baseline mean	0.788	0.788	0.785	0.785	0.790	0.790	0.788	0.788
Baseline SD	0.409	0.409	0.411	0.411	0.408	0.408	0.409	0.409
Adj. R-squared	0.353	0.354	0.320	0.319	0.387	0.392	0.355	0.357
No. observations	3190	3190	1507	1507	1683	1683	3190	3190
No. individuals	464	464	220	220	244	244	464	464
No. corrections	512	512	252	252	270	270	512	512
No. good corrections	396	396	201	201	217	217	396	396
No. bad corrections	177	177	86	86	92	92	177	177

Notes: This table presents the regression results of equation 1. 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%.

In addition, the coefficient on the female partner dummy is close to zero, suggesting that participants do not prefer male or female partners in the absence of corrections. Together with Figure 5, this indicates that participants are indifferent to a partner's gender when contributions are equal and no correction occurs.

Turning to Table 2, where I control for the partner's contribution, column 1 shows that the coefficient on any correction is large and negative: –0.202, and statistically significant at the 1% level. This implies that participants are 20.2 percentage points less willing to collaborate with someone who corrected them, regardless of whether the correction improved the puzzle outcome. A back-of-the-envelope calculation suggests that the partner's contribution would need to increase by approximately 0.84 standard deviations to offset this negative effect.[19] For male participants

[19] This number is calculated as follows: $\hat{\beta}_{\text{Partner's contribution}} \times SD_{\text{Partner's contribution}} \times x = |\hat{\beta}_{\text{Any correction}}| \Rightarrow x =$

(column 3), the corresponding coefficient is -0.171 (significant at 1%), and for female participants (column 5), it is -0.229 (also significant at 1%). Column 7 shows no statistically significant gender difference in the response to corrections. Overall, participants are less willing to collaborate with someone who corrected their move.

This negative response is problematic if participants do not distinguish between good and bad corrections. Column 2 shows that even good corrections result in a negative response: the coefficient for good correction is -0.209 and is statistically significant at the 1% level. This indicates that participants are less willing to collaborate with someone who corrected them, even when the correction was beneficial. For male participants, the corresponding coefficient is -0.181 (column 4), and for female participants, it is –0.230 (column 6); both are statistically significant at the 1% level. However, the difference between men's and women's responses is not statistically significant (column 8).

Moreover, column 2 shows that bad corrections also result in a negative response, though to a lesser extent than good corrections: the coefficient is -0.095, statistically significant at the 1% level. This effect is 0.114 percentage points smaller than the effect of good corrections, and the difference is also statistically significant at the 1% level. For male participants, the coefficient for bad corrections is statistically insignificant (column 4), whereas for female participants it is –0.153 (column 6), statistically significant at the 1% level. This suggests that women respond slightly more negatively to bad corrections (column 8, p-value < 0.10). Thus, the weaker negative response to bad corrections relative to good ones appears to be driven by male participants. However, as shown later in Figure 8, this gender difference disappears when the sample is restricted to puzzles solved within 60 seconds or to those where the partner's contribution is between 3 and 5. This may indicate that female participants are more sensitive to time pressure (Shurchkov 2012), making it harder for them to distinguish between good and bad corrections under time constraints. Alternatively, the pattern may be driven by outliers in partner contributions. Either way, the gender difference is not robust.

Taken together, these results suggest that participants are generally less willing to collaborate with partners who corrected them, even when the correction was beneficial, which reflects an irrational response. As shown in Figure 8, this pattern persists over time: limiting the analysis to earlier rounds (Rounds 1–5) does not meaningfully change the estimates.

5.2 Mechanisms

Feedback Aversion One potential mechanism is feedback aversion: a good correction may act as negative feedback about the receiving participant's ability, while a bad correction does not. The literature on information avoidance suggests that individuals with a high self-image are more averse to negative feedback (Kőszegi 2006) and are less likely to trust it (Eil and Rao 2011).

To test this hypothesis, Table 3 reports the results of equation 1, including interaction terms with a high-ability participant dummy. High-ability participants are defined as those who solved

$|\hat{\beta}_{\text{Any correction}}|/(\hat{\beta}_{\text{Partner's contribution}} \times SD_{\text{Partner's contribution}}) = 0.202/(0.084 \times 2.86) \approx 0.84$. $SD_{\text{Partner's contribution}}$ is from panel B of Table 1.

Table 3: 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		-0.165***		-0.124***		-0.213***		-0.124***
		(0.030)		(0.041)		(0.041)		(0.041)
Bad correction		-0.088*		-0.014		-0.165***		-0.014
		(0.046)		(0.062)		(0.063)		(0.062)
Any correction	-0.157***		-0.120***		-0.200***		-0.120***	
	(0.028)		(0.037)		(0.040)		(0.037)	
Female partner	0.014	0.015	0.034	0.033	-0.003	-0.004	0.034	0.033
	(0.018)	(0.018)	(0.027)	(0.027)	(0.023)	(0.023)	(0.027)	(0.027)
Partner’s contribution	0.084***	0.085***	0.078***	0.080***	0.091***	0.090***	0.078***	0.080***
	(0.003)	(0.003)	(0.004)	(0.005)	(0.004)	(0.005)	(0.004)	(0.005)
Good correction x High ability		-0.106**		-0.168**		-0.038		-0.168**
		(0.049)		(0.073)		(0.065)		(0.073)
Bad correction x High ability		-0.019		-0.052		0.029		-0.052
		(0.071)		(0.109)		(0.093)		(0.109)
Any correction x High ability	-0.106**		-0.141**		-0.062		-0.141**	
	(0.044)		(0.063)		(0.060)		(0.063)	
Female partner x High ability	-0.020	-0.020	-0.053	-0.054	0.009	0.012	-0.053	-0.054
	(0.027)	(0.027)	(0.042)	(0.043)	(0.036)	(0.036)	(0.042)	(0.043)
Partner’s contribution x High ability	-0.000	-0.000	-0.003	-0.002	-0.000	-0.000	-0.003	-0.002
	(0.005)	(0.005)	(0.007)	(0.008)	(0.007)	(0.007)	(0.007)	(0.008)
Good correction x Female								-0.088
								(0.058)
Bad correction x Female								-0.151*
								(0.088)
Any correction x Female							-0.081	
							(0.054)	
Female partner x Female							-0.037	-0.037
							(0.036)	(0.036)
Partner’s contribution x Female							0.013**	0.010
							(0.006)	(0.007)
Good correction x High ability x Female								0.130
								(0.098)
Bad correction x High ability x Female								0.082
								(0.143)
Any correction x High ability x Female							0.080	
							(0.087)	
Female partner x High ability x Female							0.062	0.065
							(0.056)	(0.056)
Partner’s contribution x High ability x Female							0.003	0.002
							(0.010)	(0.011)
Individual FE	✓	✓	✓	✓	✓	✓	✓	✓
(Good correction −Bad correction)		-0.087		-0.067		-0.116		
x High ability		(0.090)		(0.121)		(0.135)		
Good correction x High ability		-0.271***		-0.292***		-0.251***		
+Good correction		(0.039)		(0.061)		(0.050)		
Bad correction x High ability		-0.107**		-0.066		-0.136**		
+Bad correction		(0.054)		(0.089)		(0.068)		
(Good correction − Bad correction)								0.049
x High ability x Female								(0.181)
Baseline mean	0.788	0.788	0.785	0.785	0.790	0.790	0.788	0.788
Baseline SD	0.409	0.409	0.411	0.411	0.408	0.408	0.409	0.409
Adj. R-squared	0.354	0.354	0.322	0.321	0.387	0.390	0.356	0.357
No. observations	3190	3190	1507	1507	1683	1683	3190	3190
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%.

more than the median number of puzzles *in a given session* in Part 1 of the experiment (8 or more out of 15).

While high-ability participants might be expected to better distinguish between good and bad corrections, and therefore respond less negatively to both, they may also be more confident in their own abilities and thus 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. This indicates that high-ability participants are significantly less willing to collaborate with someone who corrected them, compared to low-ability participants.

Column 2 shows that this effect is primarily driven by good corrections: the interaction term for bad corrections is close to zero, while the interaction term for good corrections is negative and statistically significant at the 5% level. These results suggest that high-ability participants are particularly unwilling to collaborate with someone who corrected their mistakes, but they do not respond differently to corrections of right moves. This pattern appears to be driven by male participants (column 4) rather than female participants (column 6), although the gender difference is not statistically significant (column 8). Appendix Figure A3 confirms that these results are robust across various specifications.[20]

Corrections as Signals of Problematic Personality Another potential mechanism is that receiving a correction signals that the partner may have a difficult or unpleasant personality. The act of correcting someone could be perceived as unpleasant, making participants less inclined to collaborate with such individuals. However, this explanation does not account for the asymmetric responses to good versus bad corrections. It is also unlikely to observe such behaviors in such a short interaction.

6 Results 2: Men Respond Less Negatively to Women's Bad Corrections

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.

In column 1, the coefficient on the interaction between partner contribution and the female partner dummy is close to zero and statistically insignificant, a pattern consistent for both male participants (column 3) and female participants (column 5). As shown earlier in Figure 5, these findings suggest that participants – both men and women – neither overestimate nor underestimate women's contributions when selecting a collaborator. Similarly, the coefficient on the interaction between any correction and the female partner dummy is statistically insignificant: it is slightly

[20] 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: 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		-0.195***		-0.125**		-0.248***		-0.125**
		(0.035)		(0.053)		(0.044)		(0.053)
Bad correction		-0.182***		-0.131*		-0.213***		-0.131*
		(0.050)		(0.072)		(0.064)		(0.072)
Any correction	-0.215***		-0.155***		-0.259***		-0.155***	
	(0.031)		(0.044)		(0.042)		(0.044)	
Female partner	0.003	-0.009	0.009	-0.015	-0.000	-0.003	0.009	-0.015
	(0.021)	(0.022)	(0.029)	(0.030)	(0.030)	(0.031)	(0.029)	(0.030)
Partner's contribution	0.084***	0.083***	0.076***	0.075***	0.092***	0.091***	0.076***	0.075***
	(0.004)	(0.004)	(0.005)	(0.006)	(0.006)	(0.006)	(0.005)	(0.006)
Good correction x Female partner		-0.028		-0.103		0.033		-0.103
		(0.044)		(0.068)		(0.056)		(0.068)
Bad correction x Female partner		0.167**		0.198**		0.119		0.198**
		(0.067)		(0.098)		(0.090)		(0.098)
Any correction x Female partner	0.025		-0.028		0.059		-0.028	
	(0.040)		(0.059)		(0.055)		(0.059)	
Partner's contribution x Female partner	-0.001	0.003	0.002	0.009	-0.003	-0.002	0.002	0.009
	(0.005)	(0.005)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Good correction x Female								-0.123*
								(0.069)
Bad correction x Female								-0.082
								(0.097)
Any correction x Female							-0.104*	
							(0.061)	
Female partner x Female							-0.009	0.012
							(0.042)	(0.043)
Partner's contribution x Female							0.017**	0.017**
							(0.008)	(0.008)
Good correction x Female partner x Female								0.137
								(0.088)
Bad correction x Female partner x Female								-0.079
								(0.133)
Any correction x Female partner x Female							0.088	
							(0.080)	
Partner's contribution x Female partner x Female							-0.004	-0.010
							(0.010)	(0.010)
Individual FE	✓	✓	✓	✓	✓	✓	✓	✓
(Good correction − Bad correction)		-0.195**		-0.301**		-0.085		
x Female partner		(0.085)		(0.129)		(0.111)		
Good correction x Female partner		-0.224***		-0.228***		-0.215***		
+Good correction		(0.030)		(0.044)		(0.041)		
Bad correction x Female partner		-0.015		0.067		-0.095		
+Bad correction		(0.046)		(0.065)		(0.063)		
(Good correction − Bad correction)								-0.216
x Female partner x Female								(0.170)
Baseline mean	0.788	0.788	0.785	0.785	0.790	0.790	0.788	0.788
Baseline SD	0.409	0.409	0.411	0.411	0.408	0.408	0.409	0.409
Adj. R-squared	0.352	0.355	0.319	0.320	0.387	0.392	0.355	0.358
No. observations	3190	3190	1507	1507	1683	1683	3190	3190
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%.

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 separate good and bad corrections, as shown 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 at the 5% level. Moreover, the coefficient estimates for women's good and bad corrections differ significantly at the 5% level. This indicates that participants are less willing to collaborate with a female partner who corrected their mistakes than with a female partner who corrected their right moves. These patterns are driven primarily by male participants (column 4), with no significant effects observed among female participants (column 6). A caveat is that the difference between male and female participants is not statistically significant, so this evidence should be interpreted as suggestive.

Taken together, these results suggest that men are more tolerant of women's bad corrections but are just as intolerant of good corrections from women as they are of those from men. Women do not exhibit this asymmetry: they respond equally to women's and men's good and bad corrections. This can result in gendered sorting into teams, with mixed-gender teams attracting less competent women than women-only teams.

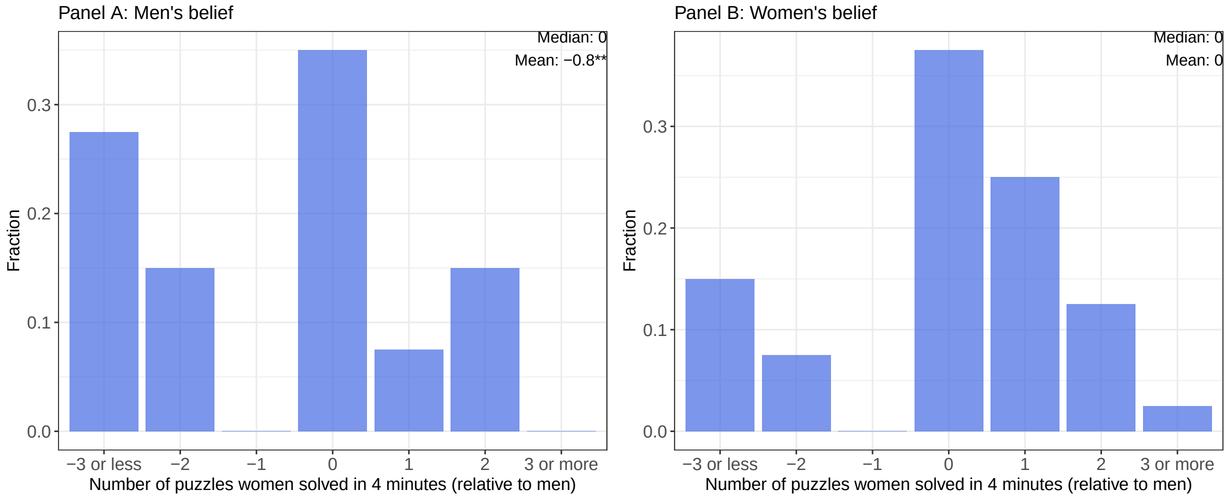
6.2 Mechanisms

Gender Bias One potential explanation for men's asymmetric reactions to women's corrections is gender bias: participants who hold negative biases against women might respond differently to good and bad corrections from female partners. However, this explanation is not supported by the data. Appendix Table A3 reports the results of equation 1 with interaction terms for a high gender bias dummy. Participants are classified as having high gender bias if their score on six hostile and benevolent sexism items from Karpowitz et al. (2024) is above the median for their own gender.

Across all specifications – whether using the full sample (column 2), male participants only (column 4), or female participants only (column 6) – the interaction between good correction and high gender bias is statistically insignificant. The same is true for the interaction between bad correction and high gender bias. These results suggest that general gender bias does not account for the differential responses to women's corrections. Thus, it is not gender bias – that is, not a general dislike of being corrected by women – that drives men's asymmetric response.

Belief about Women's Puzzle-Solving Ability Another potential mechanism is participants' beliefs about gender differences in puzzle-solving ability. While participants exhibit no baseline preference between male and female collaborators when contributions are held constant (as shown in Figure 5 and Table 2), stereotypes about women's abilities may lead to asymmetric reactions once corrections occur. Specifically, if participants believe women are less competent at solving puzzles, then bad corrections from women may be perceived as less ego-threatening than good corrections from them, leading to differential reactions. While this may seem contradictory to the lack of baseline preference differences, it is consistent with motivated stereotyping; for example, Sinclair and Kunda (2000) show that individuals express stereotypes only when it helps protect their egos.

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 supports this interpretation. Panel A shows that men believe women are slightly less capable puzzle-solvers than men, estimating that women solved 0.8 fewer puzzles on average in Part 1 of the main experiment. In contrast, Panel B shows that women do not perceive a significant gender gap in puzzle-solving ability. These belief patterns are consistent with the behavioral results: men, but not women, react differently to good versus bad corrections from female partners.

7 Robustness Checks

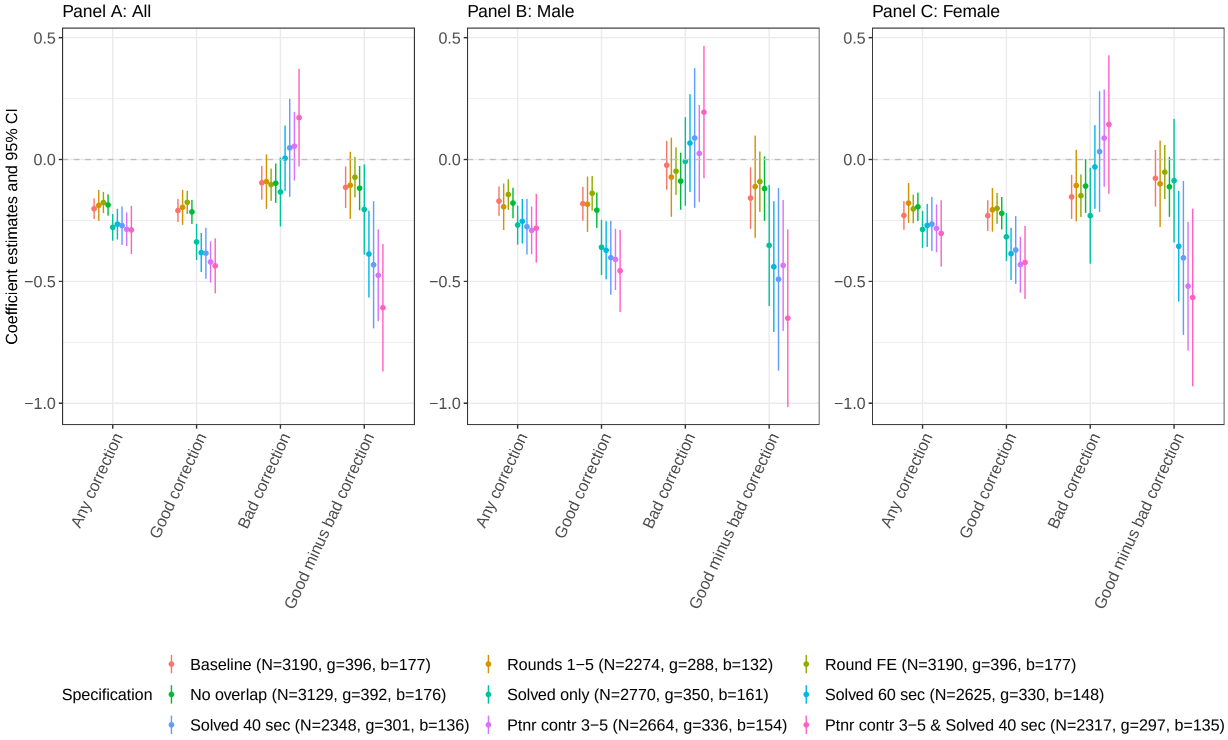
7.1 Robustness Checks for Results 1

Across-Round Imbalance As shown in Figure 6, participants in rounds 6 and 7 were less likely to collaborate, more likely to correct their partners, and less likely to solve the puzzle-indicating some imbalance across rounds. To address this, I re-estimate the coefficients for good and bad corrections, as well as their difference, excluding rounds 6 and 7 (labeled "Rounds 1-5"), and with round fixed effects ("Round FE"), as reported in Figure 8.

Panel A shows that the estimates for all participants remain consistent with the baseline specification, though the confidence intervals widen slightly due to the smaller sample size. Panels B and C show similar patterns for male and female participants, respectively. These results suggest that the main findings are not driven by across-round imbalances.

Puzzles with Both Good and Bad Corrections As discussed in Section 3.1, 61 puzzles included both good and bad corrections, which could introduce ambiguity regarding which correction affected participants' collaboration decisions. To address this concern, I re-estimate the regression excluding these puzzles ("No overlap" in Figure 8). The results remain robust and are nearly

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 all participants, Panel B for male participants only, and Panel C for female participants only. Sample sizes (N), good corrections (g), and bad corrections (b) for each specification are shown in parentheses.

identical to the baseline estimates across all subgroups, confirming that overlapping corrections do not drive the results.

Unsolved Puzzles As discussed in Section 3.1, 13% of puzzles were unsolved. To ensure that these puzzles are not driving the results, I re-estimate the coefficients for good and bad corrections using only solved puzzles ("Solved only" in Figure 8). The coefficient for good corrections becomes 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-estimate the coefficients for puzzles solved within 60 seconds ("Solved 60 sec") and 40 seconds ("Solved 40 sec") in Figure 8. The coefficient for any correction becomes slightly more negative than in the baseline. The coefficient for good corrections becomes more negative as well, while the coefficient for bad corrections approaches zero. These findings suggest that solution time does not mediate the results and further support the conclusion that participants do not respond rationally to corrections.

Outliers and Functional Form of Partner Contribution As shown in Figure 4, most partner contributions fall between 3 and 5, though some outliers exist. Additionally, the empirical strategy assumes a linear relationship between partner contributions and perceived ability, which may not fully capture the effect. To address these concerns, I re-estimate the coefficients using puzzles where partner contributions fall between 3 and 5 ("Ptnr contr 3-5"), and those within this range that were also solved within 40 seconds ("Ptnr contr 3-5 & Solved 40 sec"). The estimates remain consistent with the baseline, and this holds for both male (Panel B) and female participants (Panel C), suggesting that outliers do not drive the results. These robustness checks indicate that it is the act of being corrected, rather than the correction's quality, that reduces participants' willingness to collaborate, which constitutes an irrational response.

Finally, the number of good and bad corrections, listed in parentheses next to each specification's name, is sufficiently large for every specification. Thus, the (in)significance of the coefficient estimates cannot be attributed to an insufficient number of corrections.

7.2 Robustness Checks for Results 2

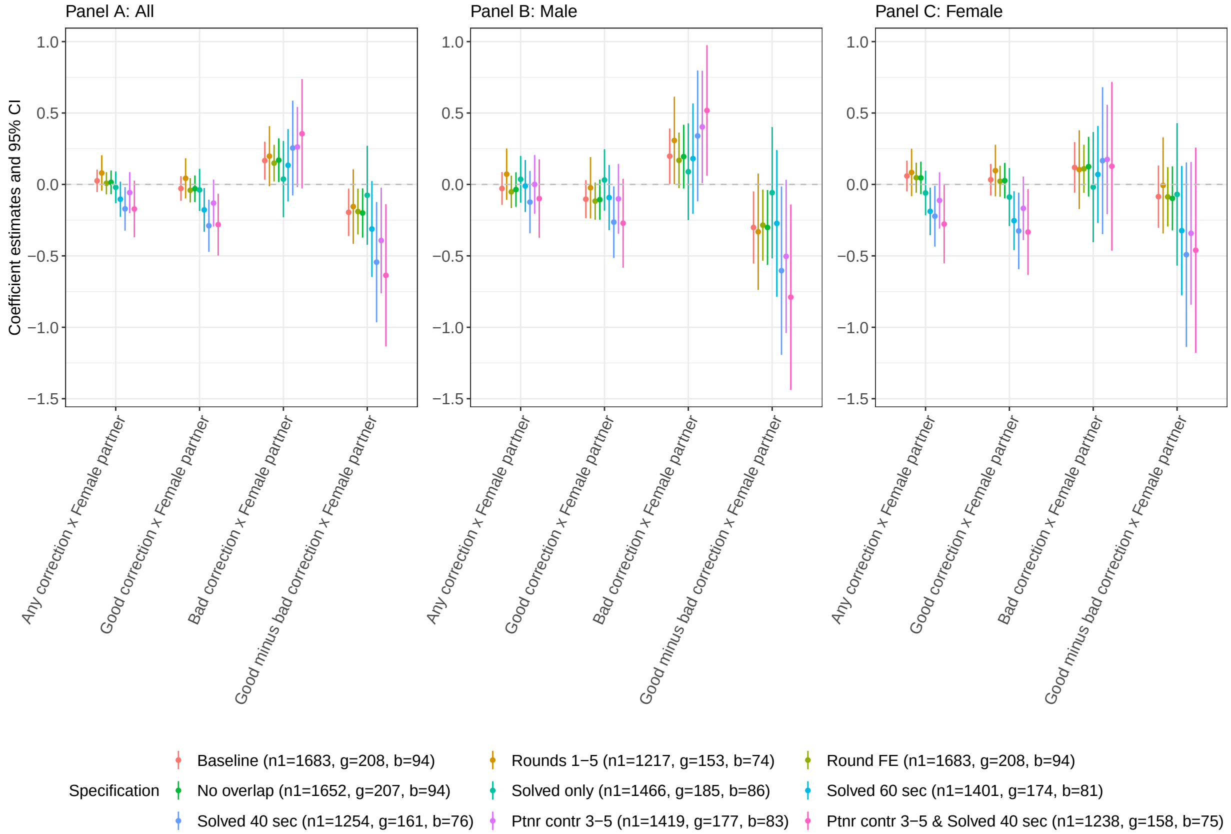
Figure 9 presents the same robustness checks as Figure 8, confirming that the results reported in Table 4 are generally robust across various specifications. However, one notable exception emerges: when the sample is restricted to puzzles solved within 60 seconds, the negative effect of female partners' good corrections becomes statistically significant at the 5% level, and this effect is primarily driven by female participants. Yet, the overall pattern remains: the asymmetric response to female partners' good versus bad corrections is predominantly observed among male participants. For female participants, the asymmetric response is not statistically significant in any of the specifications.

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

8 Conclusion

Teamwork is increasingly essential in modern workplaces, yet interpersonal frictions can undermine its effectiveness. This paper highlights an important barrier to successful collaboration: individuals' reluctance to work with those who corrected them. I show that individuals are significantly less willing to collaborate with someone who corrected them, even when the correction benefits the team. The likely mechanism is negative feedback aversion: participants who received positive feedback about their ability were much less willing to collaborate with someone who made good corrections that benefit the team, but not with those who made bad corrections that hinder the team's progress. Furthermore, I find that men react less negatively to women's bad corrections but equally negatively to women's good corrections, potentially due to their beliefs about women's abilities. Women, on the

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 shows estimates for all participants, Panel B for male participants only, and Panel C for female participants only. Number of observations (n1), good corrections (g), and bad corrections (b) where the partner is female are indicated in parentheses.

other hand, respond women’s and men’s corrections equally negatively. This can induce gendered sorting into teams, with mixed-gender teams attracting less competent women.

Of course, the effect sizes observed in this study may differ in real-world contexts. Several workplace dynamics absent from my experiment could amplify negative reactions to corrections, including: (i) reputational costs (Bénabou and Tirole 2006), (ii) emotional stakes in tasks, and (iii) hierarchical structures. First, the emotional cost of being corrected is likely higher when others are present. Second, workplace tasks often carry more personal significance than the experimental puzzle, making criticism feel more consequential. Third, since all participants in my study were equals, it did not capture the dynamics of junior-senior relationships, where corrections from juniors may provoke stronger negative reactions.

Conversely, certain real-world factors could mitigate negative responses to corrections. These include (i) pre-existing relationships and (ii) ambiguity in communication. In the experiment, participants were strangers, whereas workplace collaborators typically know each other. A correction from a trusted colleague may be received more favorably, or more harshly if the relationship is strained. Additionally, corrections in the workplace are often subtler and less binary than in the experimental setting. These considerations suggest that fostering strong interpersonal relationships may help buffer negative reactions to corrections and support effective teamwork.

This paper offers a controlled experimental benchmark for understanding how corrections influence team dynamics. The findings underscore the importance of interpersonal factors in collaboration and how women and men are treated differently in the team, and highlight the potential for future research in more complex and hierarchical environments.

References

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.

Azmat, Ghazala, Manuel Bagues, Antonio Cabrales, and Nagore Iriberri. 2019. “What You Don’t Know…Can’t Hurt You? A Natural Field Experiment on Relative Performance Feedback in Higher Education.” *Management Science* 65 (8): 3714–3736.

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.

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.

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.*

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

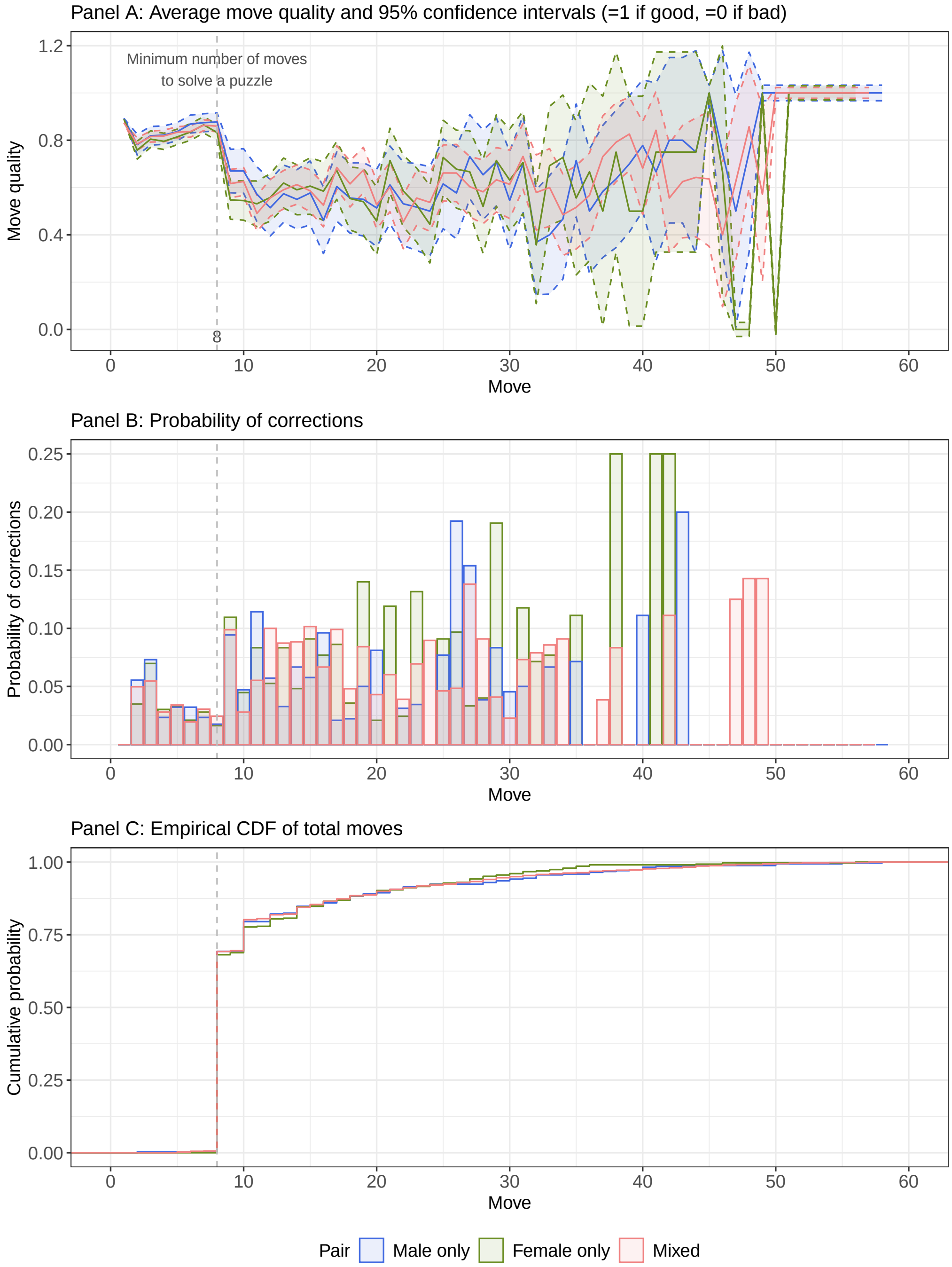
A Additional Figures and Tables

Table A1: Participants' Characteristics

	All (N=464)		Male (N=220)		Female (N=244)		Difference (Male – Female)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SE
Age	25.12	3.81	25.87	4.33	24.45	3.13	1.41***	0.35
Gender bias [0-1]	0.22	0.19	0.29	0.19	0.17	0.16	0.12***	0.02
Region of origin (within Italy)								
North	0.34		0.36		0.32		0.04	0.04
Center	0.23		0.24		0.23		0.01	0.04
South	0.42		0.40		0.45		-0.06	0.05
Major:								
Humanities	0.34		0.22		0.45		-0.23***	0.04
Social sciences	0.25		0.27		0.24		0.03	0.04
Natural sciences	0.16		0.20		0.12		0.08**	0.03
Engineering	0.14		0.23		0.05		0.17***	0.03
Medicine	0.11		0.08		0.13		-0.05*	0.03
Program:								
Bachelor	0.31		0.26		0.34		-0.08*	0.04
Master	0.65		0.68		0.63		0.05	0.04
Doctor	0.04		0.06		0.03		0.03	0.02

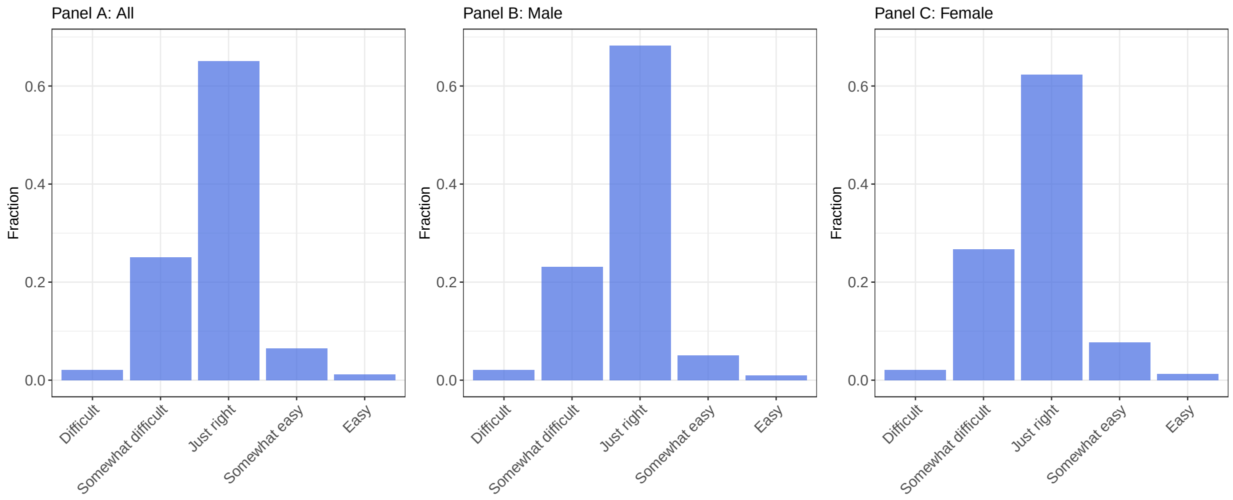
Notes: This table describes participants' characteristics. P-values of the difference between male and female participants are calculated with heteroskedasticity-robust standard errors. Significance levels: * 10%, ** 5%, and *** 1%.

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



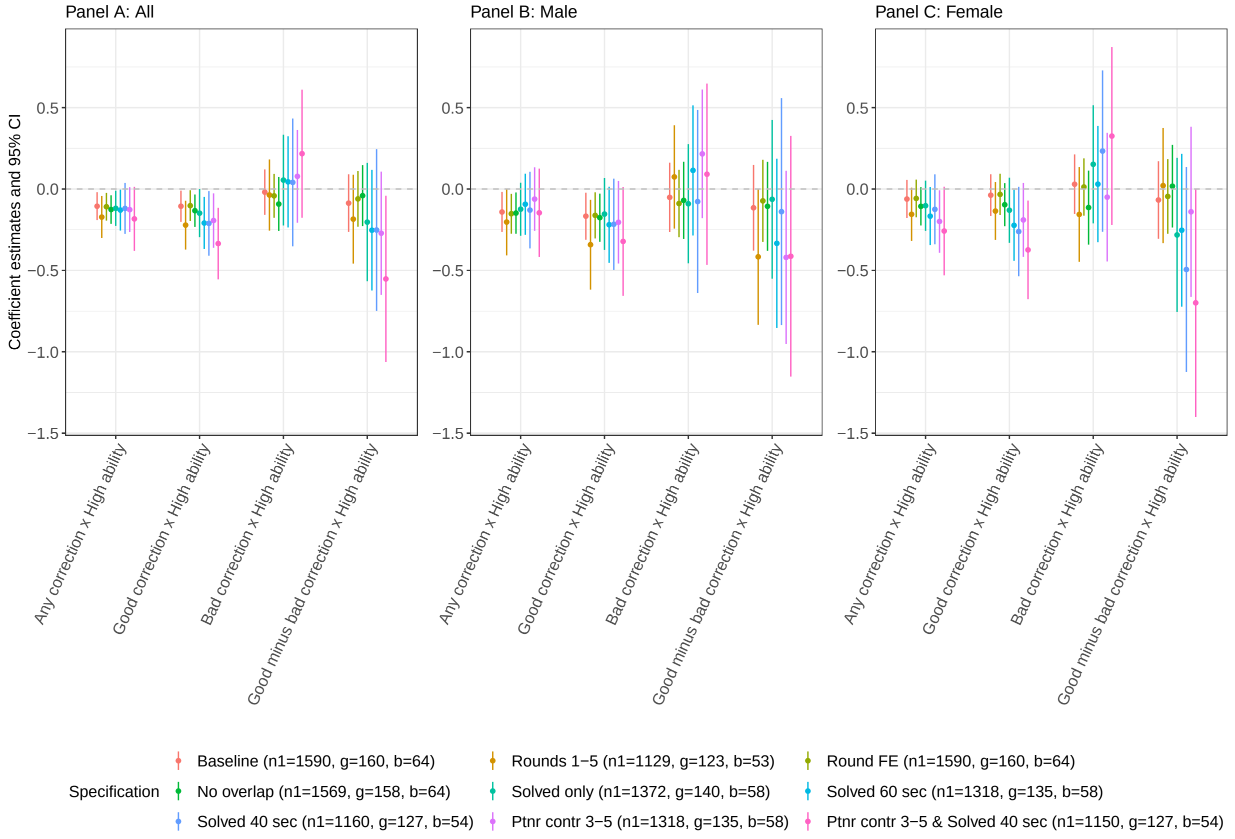
Notes: This figure shows the participants' perceived puzzle difficulty from the post-questionnaire. Panel A shows the perception of all participants, Panel B shows the perception of male participants, and Panel C shows the perception of female participants.

Table A2: Effect of Receiving Corrections on Willingness to Collaborate (without controlling for partner contribution)

Dependent variable:	Willing to collaborate (yes=1, no=0)							
Sample:	All		Male		Female		All	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Good correction	-0.222***	-0.246***	-0.194***	-0.221***	-0.245***	-0.266***	-0.203***	-0.221***
	(0.027)	(0.029)	(0.040)	(0.039)	(0.037)	(0.042)	(0.040)	(0.039)
Bad correction	-0.525***	-0.513***	-0.484***	-0.463***	-0.558***	-0.554***	-0.491***	-0.463***
	(0.030)	(0.033)	(0.048)	(0.051)	(0.038)	(0.043)	(0.048)	(0.051)
Female partner	-0.010	-0.011	0.006	-0.002	-0.024	-0.018	-0.013	-0.002
	(0.016)	(0.017)	(0.024)	(0.026)	(0.020)	(0.022)	(0.019)	(0.026)
Good correction x Female							-0.034	-0.045
							(0.053)	(0.058)
Bad correction x Female							-0.060	-0.092
							(0.061)	(0.067)
Female partner x Female							0.004	-0.015
							(0.021)	(0.034)
Individual FE		✓		✓		✓		✓
Good correction	0.303***	0.267***	0.291***	0.242***	0.313***	0.288***		
−Bad correction	(0.047)	(0.051)	(0.070)	(0.072)	(0.065)	(0.073)		
Baseline mean	0.788	0.788	0.785	0.785	0.790	0.790	0.788	0.788
Baseline SD	0.409	0.409	0.411	0.411	0.408	0.408	0.409	0.409
Adj. R-squared	0.113	0.106	0.085	0.100	0.139	0.113	0.113	0.106
No. observations	3190	3190	1507	1507	1683	1683	3190	3190
No. individuals	464	464	220	220	244	244	464	464

Notes: This table presents the regression results of equation 1 without controlling for partner contribution. 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%.

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 A3: 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		-0.190***		-0.087		-0.264***
		(0.046)		(0.070)		(0.057)
Bad correction		-0.204***		-0.219**		-0.186**
		(0.071)		(0.102)		(0.093)
Any correction	-0.227***		-0.161**		-0.269***	
	(0.041)		(0.063)		(0.052)	
Female partner	0.012	-0.007	0.017	-0.014	0.009	0.001
	(0.028)	(0.029)	(0.042)	(0.044)	(0.037)	(0.038)
Partner's contribution	0.088***	0.087***	0.081***	0.078***	0.095***	0.095***
	(0.005)	(0.006)	(0.008)	(0.009)	(0.007)	(0.007)
Good correction x Female partner		-0.024		-0.150*		0.068
		(0.059)		(0.086)		(0.077)
Bad correction x Female partner		0.259***		0.338**		0.194
		(0.094)		(0.141)		(0.123)
Any correction x Female partner	0.046		-0.041		0.105	
	(0.054)		(0.077)		(0.074)	
Partner's contribution x Female partner	-0.003	0.002	-0.001	0.009	-0.005	-0.003
	(0.007)	(0.007)	(0.010)	(0.011)	(0.009)	(0.010)
Good correction x High gender bias		-0.016		-0.075		0.035
		(0.071)		(0.106)		(0.092)
Bad correction x High gender bias		0.037		0.140		-0.058
		(0.098)		(0.138)		(0.129)
Any correction x High gender bias	0.018		0.003		0.021	
	(0.062)		(0.089)		(0.088)	
Female partner x High gender bias	-0.025	-0.013	-0.031	-0.013	-0.020	-0.012
	(0.043)	(0.044)	(0.058)	(0.059)	(0.061)	(0.062)
Partner's contribution x High gender bias	-0.009	-0.009	-0.012	-0.008	-0.006	-0.008
	(0.008)	(0.009)	(0.010)	(0.011)	(0.012)	(0.012)
Good correction x Female partner x High gender bias		-0.001		0.091		-0.061
		(0.089)		(0.136)		(0.115)
Bad correction x Female partner x High gender bias		-0.170		-0.235		-0.144
		(0.133)		(0.191)		(0.179)
Any correction x Female partner x High gender bias	-0.037		0.031		-0.090	
	(0.081)		(0.117)		(0.112)	
Partner's contribution x Female partner x High gender bias	0.006	0.003	0.007	0.002	0.005	0.003
	(0.010)	(0.011)	(0.013)	(0.014)	(0.015)	(0.015)
Individual FE	✓	✓	✓	✓	✓	✓
Baseline mean	0.789	0.789	0.785	0.785	0.790	0.790
Baseline SD	0.408	0.408	0.411	0.411	0.408	0.408
Adj. R-squared	0.352	0.354	0.318	0.319	0.386	0.391
No. observations	3183	3183	1500	1500	1683	1683
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

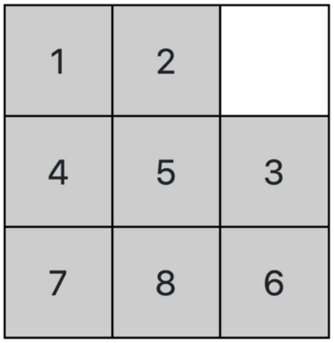
[I was chosen to participate] [I was chosen to leave]

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.



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.

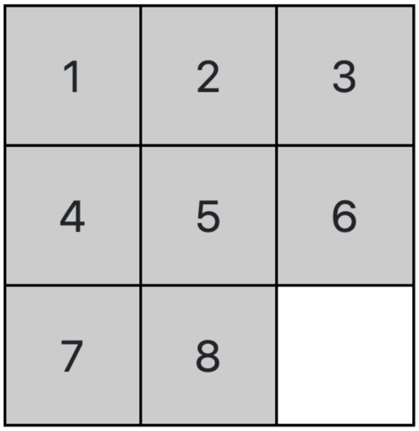
App: pt1

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:



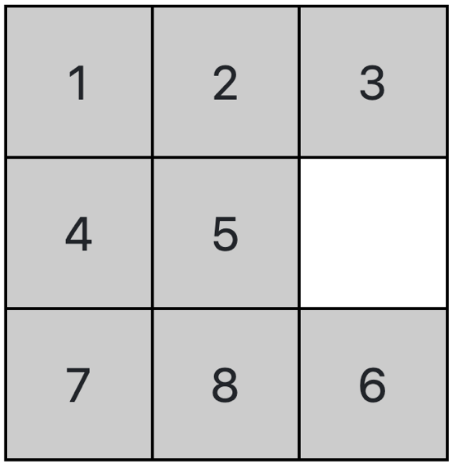
1	2	3
4	5	6
7	8	

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

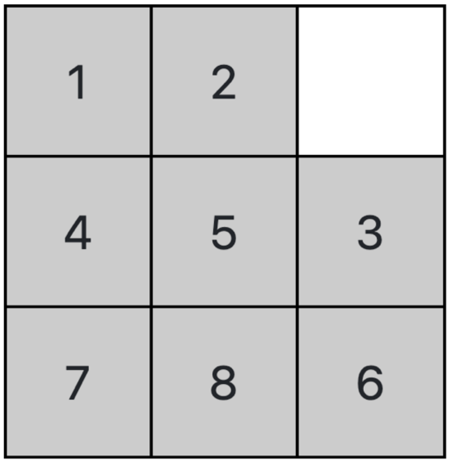
You can only move the tiles next to an empty cell and the tile you choose is moved to the empty cell. So, in this puzzle, there are 3 moves you can make: move 3 down, move 5 right, and move 6 up.

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

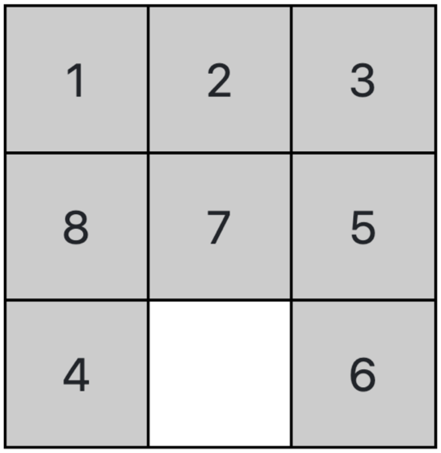
First, there are 2 moves you can make: move 2 right and move 3 up. Which moves should you make?

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>:

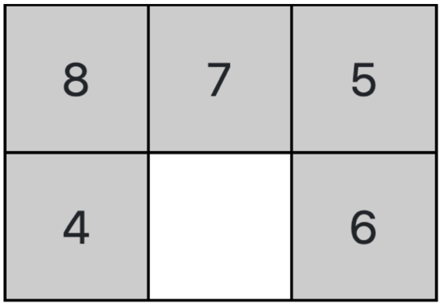
Finally, consider the following puzzle.



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.

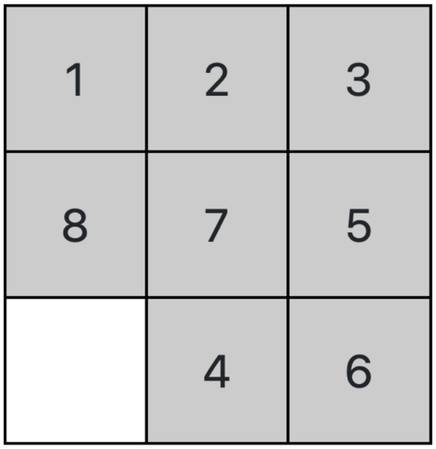


8	7	5
4		6

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?

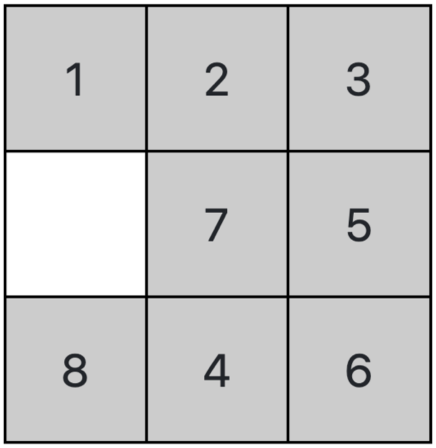
Let's try moving 4 right.



1	2	3
8	7	5
	4	6

puzzle_3_1_bad_0.png

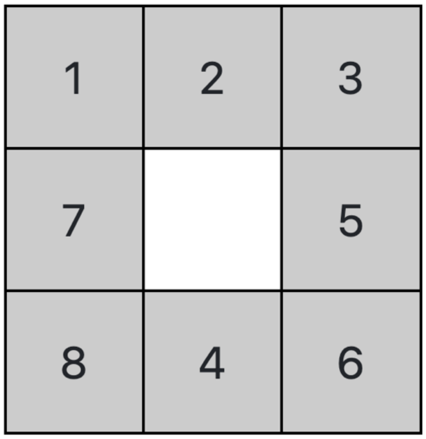
Now the only tile you can move is 8. So, let's move it down.



1	2	3
	7	5
8	4	6

puzzle_3_1_bad_1.png

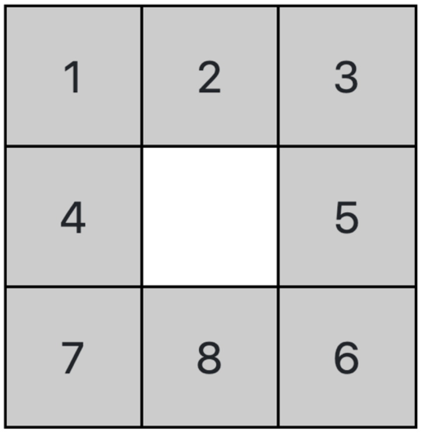
Now, if you ignore the top row which is already in the correct order, the only tile you can move is 7. So, let's move it to the left.



1	2	3
7		5
8	4	6

puzzle_3_1_bad_2.png

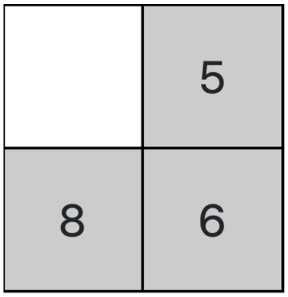
Then move 4 up, move 8 right, and move 7 down. Then you have made the left column in the correct order. You have moved tiles seven times until now.



1	2	3
4		5
7	8	6

puzzle_3_1_bad_3.png

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

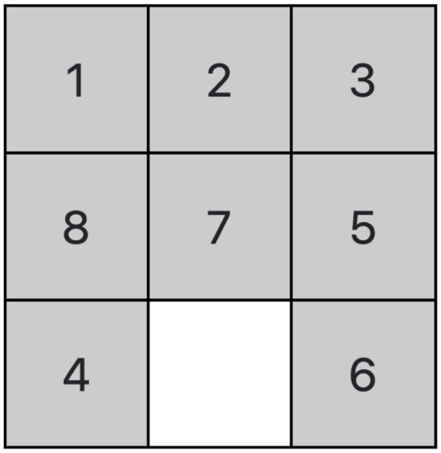


	5
8	6

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**.

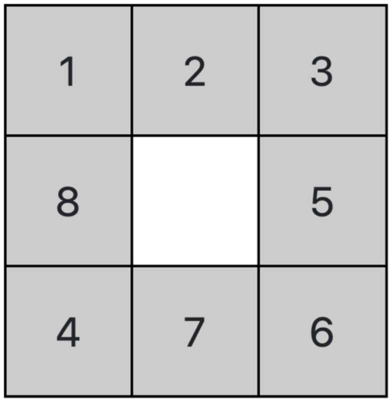
Let’s go back to the initial puzzle.



1	2	3
8	7	5
4		6

puzzle_3_0.png

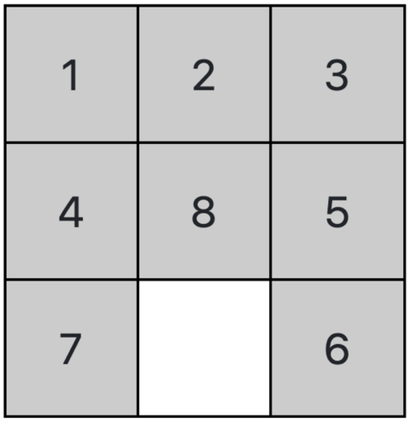
This time, let’s try moving 7 down.



1	2	3
8		5
4	7	6

puzzle_3_1_good.png

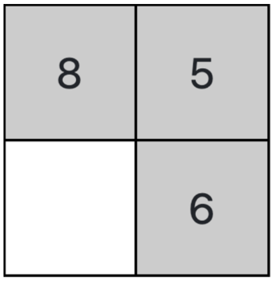
Then move 8 right, 4 up, and 7 left. Now you have made the left column in the correct order only with four moves.



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.

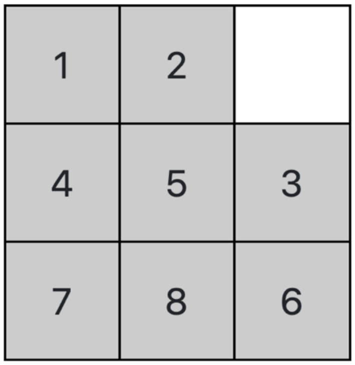
1. Which of the following statements is true?

- ✓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

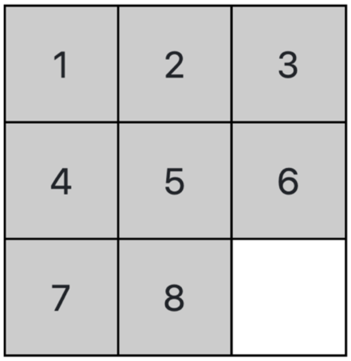
A



1	2	
4	5	3
7	8	6

puzzle_2_0.png

B



1	2	3
4	5	6
7	8	

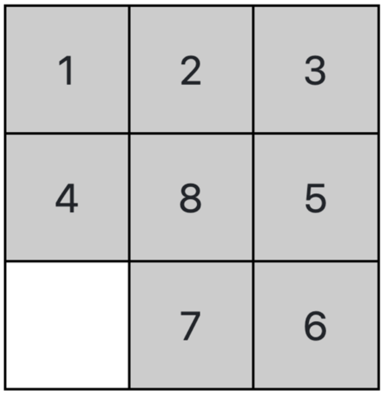
puzzle_goal.png

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

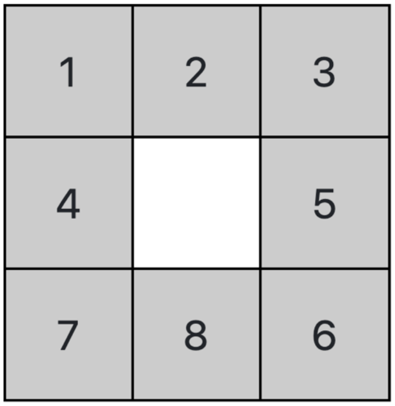
puzzle_3_3_good.png

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.

<u>Understanding questions</u>:

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

1. Which of the following statements is true?

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

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

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

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

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

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

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

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

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

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

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

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

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

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

- Giovanni
- Valeria

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

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

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

Page: Puzzle

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

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

[2 minutes time count]

Page: Pref

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

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

[Yes, No]

App: pt3

Page: Partner

Your partner for part 3

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

Page: Intro

Instructions for part 3 out of 3

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

Understanding questions:

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

1. Which of the following statements is true?

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

2. You and your partner…

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

Page: Ready

Be ready

[5 seconds time count]

Please be ready for the group round.

Page: Game

Puzzle 1/2/3/…/20

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

[12 minutes time count]

[max. 20 puzzles with increasing difficulty]

Page: Proceed

The group round is over

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

Please click Next to proceed.

<u>App: pt4</u>

Page: Intro

A follow-up survey

As the last task, we will ask you a series of questions in which there are no right or wrong answers. We are only interested in your personal opinions. We are interested in what characteristics are associated with people's behaviors in this study. **The answers you provide will in no way affect your earnings in this study and are kept confidential.**

Please click Next to start the survey.

Page: SurveyASI

Survey page 1 out of 2

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

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

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

Page: SurveyDem

Survey page 2 out of 2

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

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

Page: ThankYou

Thank you for your participation

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

Your earnings:

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

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

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

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

App: pt99

Page: ThankYou

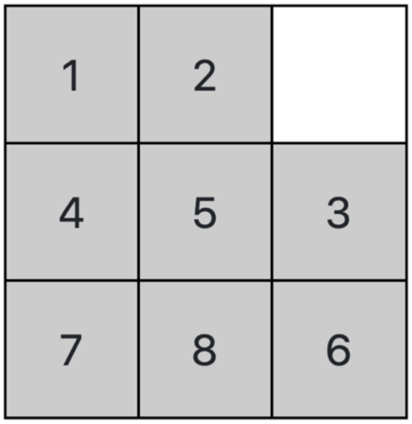
Thank you for showing up

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

Welcome!

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

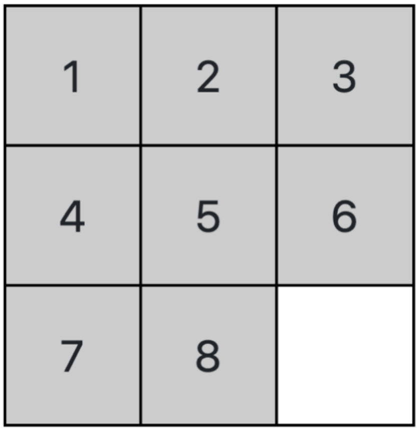
In this study, you will solve one puzzle, like the one shown below. After you solve the puzzle, we will ask you to guess about the puzzle. If your guess is correct, you will receive a bonus payment of £1.



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.

Comprehension questions:

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

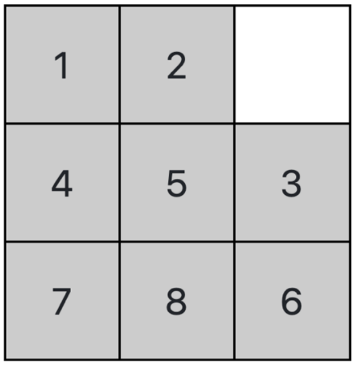
2. Which tiles can you move?

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

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

- A
- B

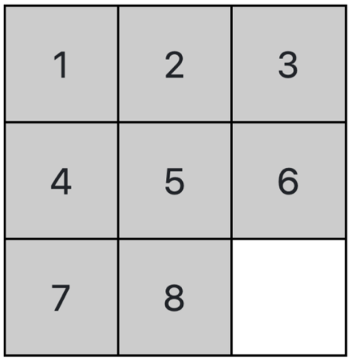
A



1	2	
4	5	3
7	8	6

puzzle_2_0.png

B



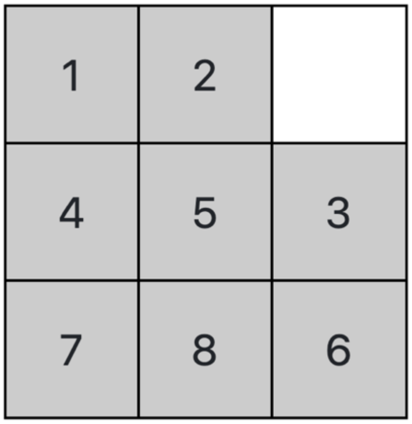
1	2	3
4	5	6
7	8	

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.