Corrections and Collaborations in Group Work

Yuki Takahashi*

Preliminary draft. Please do not cite or circulate.

Click here for the latest version

June 11, 2021

Abstract

Receiving corrections from colleagues is an integral part of group work. However, people may take them emotionally, which could be very costly for a person who corrects them as collaboration is essential in group work. This paper studies how being corrected by others in a group affects one's willingness to collaborate with those people in later works in a quasi-laboratory experimental setting. I find that the main determinant of collaborator selection is a given person's contribution to the task. However, after controlling for the contribution, people are significantly less willing to collaborate with a person who has corrected their actions. Women do not like being corrected both for their mistakes and for their right actions, while men mostly do not like being corrected for their mistakes. High-ability men especially do not like to be corrected for their mistakes, suggesting that their emotional irritation is driving their negative reactions. The gender of the person who made corrections does not matter. These findings have implications for organizational efficiency, gender differences in managerial practice, corrections, and strategic behaviors.

JEL codes: D91, C92, M54, J16
Keywords: correction, collaboration, group work, gender differences

---

*PhD Candidate, Department of Economics, University of Bologna. Email: yuki.takahashi2@unibo.it. I am grateful to Maria Bigoni, Siri Isaksson, Bertil Tungodden, Laura Anderlucci, and Natalia Montinari whose feedback was essential for this project. I am also grateful to participants of the experiment for their participation and cooperation. Francesca Cassanelli, Alessandro Castagnetti, Mónica Costa-Dias, Ria Granzier-Nakajima, Silvia Griselda, Annalisa Loviglio, Øivind Schøyen, Vincenzo Scrutinio, Erik Ø. Sørensen, Ludovica Spinola, and PhD students at the NHH and the University of Bologna all provided many helpful comments. This paper also benefited from participants' comments at the Annual Southern PhD Economics Conference, the PhD-EVS, the WEAI, and seminars at Ca' Foscari University, Catholic University of Brasília, the NHH, the University of Bologna, and the University of Copenhagen. Ceren Ay, Tommaso Batistoni, Philipp Chapkovski, Sebastian Fest, Christian König genannt Kersting, and oTree help & discussion group kindly answered my questions about oTree programming; in particular, my puzzle code was heavily based on Christian's code. Michela Boldrini and Boon Han Koh conducted the quasi-laboratory experiments ahead of me and kindly answered my questions about their implementations. Lorenzo Golinelli provided excellent technical and administrative assistance. This study was pre-registered with the OSF registry (https://osf.io/tgyc5) and approved by the IRB at the University of Bologna (#262643).

1 Introduction

Receiving corrections from colleagues is an integral part of group work. Consider academic research. From the development of ideas to the writing up of the final draft, we discuss our research project with several colleagues, receive criticisms from them, and refine the ideas and the analysis.

However, being corrected by others could make us emotionally irritated rather than appreciating the corrections, e.g., imagine you present a paper for which you spent several years and someone points out a failure of your identification assumption. This emotional reaction could be very costly for a person who corrects others because she may reduce the probability of having a collaboration, which could be important for her career success.[1] Consider two assistant professors, one with several collaboration works with her colleagues and the other without collaboration works. When they face tenure evaluation, the former assistant professor is more likely to have a better publication record than the latter assistant professor and more likely to get tenure.

This paper studies how being corrected by others in a group affects one's willingness to collaborate with those people in later works. To answer this question, I design a quasi-laboratory experiment, a hybrid of physical laboratory and online experiments. In the experiment, participants are grouped into people of eight, paired with another group member, solve one joint task together by alternating their moves. After solving the task, participants state whether they would like to collaborate with the same group member for the same task in the next stage, which is the main source of earnings. This gives a strong incentive for participants to select as good a collaborator as possible. Participants are paired with all the seven group members in a random order to address endogenous group formation. As a joint task, I use Isaksson (2018)'s number-sliding puzzle which allows me to calculate an objective measure of each participant's contribution to the joint task as well as to classify each move as good or bad.[2] I define a correction as reversing a group member's move, which is comparable across different participants.

I find that the main determinant of participants' collaborator selection is paired participants' contribution to the puzzle. However, after controlling for the contribution, people are significantly less willing to collaborate with a paired participant who has corrected their moves. Women react negatively to corrections of their mistakes and their right moves, while men react negatively to corrections of their mistakes. High-ability men react particularly negatively to corrections of their mistakes, suggesting that it is not their Bayesian updating but their emotion that is driving the results. The gender of the person who makes corrections does not matter for people's negative reactions.

These findings have implications for three strands of literature. The first strand of literature is organizational efficiency. The literature finds that managers often favor workers whom they like in compensation and promotion (MacLeod 2003; Prendergast and Topel 1996) and workers tend to conform their managers (Prendergast 1993), both of which distorts the optimal allocation of talent.[3] In addition, Li (2020) finds that this managers' favoritism not only distorts the optimal

1. In economics, for example, most papers are co-authored (Jones 2021).
2. Participants solve a 3x3 number-sliding puzzle in pairs by alternating their moves.
3. Several studies empirically verify these theoretical findings in various different settings (Bandiera, Barankay,

allocation of talent but also reduces non-favored workers' performance. Also, the literature finds that people tend to view others who disagree with them as biased (Kennedy and Pronin 2008) and as having immoral motives (Reifen Tagar 2014). My findings that people's reluctance to be corrected can be another source of distortion.

My findings also have implications for literature on gender differences in managerial practice. The literature finds that female teams employ less aggressive strategy (Apesteguia, Azmat, and Iriberri 2011), female leaders take less risk for a group (Ertac and Gurdal 2012), less likely to exaggerate their past performance (Reuben et al. 2012), and less likely to make assertive cheap talk (Manian and Sheth 2021). Matsa and Miller (2013) find that firms with a larger fraction of female board members undertake fewer worker layoffs. My findings that men are reluctant to being corrected for their mistakes but not for their right actions could discourage workers to correct their male leaders.

In addition, my findings have implications for literature on gender differences in corrections and strategic behaviors. The literature finds that group members correct women's ideas more often than men's ideas (Guo and Recalde 2020) and that men are more likely to correct their group member's bad moves in the same puzzle used in my experiment (Isaksson 2018).[4] The literature also finds that women make higher offer in ultimatum games and female pairs are much more likely to sustain cooperation (Eckel and Grossman 2001),[5] and women retaliate more strategically than men (Dehdari, Heikensten, and Isaksson 2019). I enrich this literature by showing gender differences in response to being corrected.

2 Experiment

There are two main empirical challenges to examine the effect of corrections on collaborator selections using secondary data. First, group formation is not random but corrections are endogenous to the group structure. Second, different corrections are not necessarily comparable to each other. Thus, I test my question in a controlled quasi-laboratory experimental setting where group formation is randomized and define corrections in a puzzle where researchers can track mathematically whether a given correction helped or did not help to solve the puzzle.

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

and Rasul 2009; Beaman and Magruder 2012; Hjort 2014; Xu 2018).

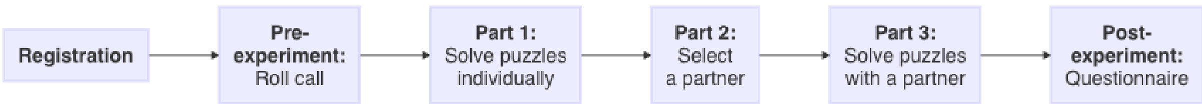
4. As the puzzle was originally used by Isaksson (2018).

5. Solnick (2001) does not find that women make higher offer; Croson and Gneezy (2009) discuss potential reasons for the different results. In dictator games, the literature finds that women give more in dictator games (Eckel and Grossman 1998) and prefer equal splits more than men in modified dictator games (Andreoni and Vesterlund 2001).

Group task As the group task I use Isaksson (2018)'s puzzle, a sliding puzzle with 8 numbered tiles, which should be placed in numerical order within a 3x3 frame (see figure 3 for an example). To achieve this goal, participants play in pairs, alternating their moves. This puzzle has nice mathematical properties that I can define the puzzle difficulty and one's good and bad moves by the Breadth-First Search algorithm, from which I can calculate individual contributions to the group task and the quality of corrections objectively and comparably.[6] Further, the puzzle-solving captures an essential characteristic of group work in which two or more people work towards the same goal (Isaksson 2018) but the quality of each move and correction is only partially observable to participants (but fully observable to the experimenter).

The experiment consists of three parts as summarized in figure 1 and described in detail below. At the beginning of each part, participants must answer a set of comprehension questions to make sure they understand the instructions.

Figure 1: Flowchart of the experiment



Notes: This figure shows an overview of the experiment discussed in detail in section 2.1.

2.1 Design and procedure

Registration

Upon receiving an invitation email to the experiment, participants register for a session they want to participate in and upload their ID documents as well as a signed consent form.[7]

Pre-experiment

On the day and the time of the session they have registered for, participants enter the Zoom waiting room.[8] They receive a link to the virtual room for the experiment and enter their first name, last name, and their email they have used in the registration. They also draw a virtual coin numbered from 1 to 40 without replacement.

Then I admit participants to the Zoom meeting room one by one and rename them by the first name they have just entered. If there is more than one participant with the same first name, I add a number after their first name (e.g. Giovanni2).

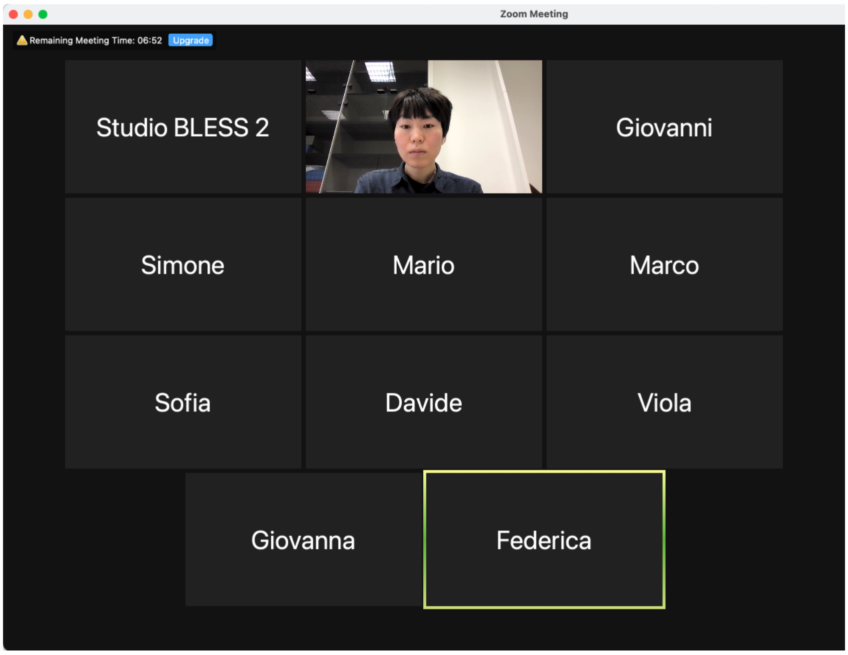
After admitting all the participants to the Zoom meeting room, I do roll call (Bordalo et al. 2019; Coffman, Flikkema, and Shurchkov 2021): I take attendance by calling each participant's first name one by one and ask her or him to respond via microphone. This process ensures other participants that the called participant's first name corresponds to her or his gender. If there

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

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

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

Figure 2: Zoom screen



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

are more participants than I would need for the session (I need 16 participants), I draw random numbers from 1 to 40 and ask those who drew the coins with the same number to leave.[9] Those who leave the session receive the 2€ show-up fee. Figure 2 shows a Zoom screen participants would see during the roll call (the person whose camera is on is the experimenter; participants would see this screen throughout the experiment but the experimenter's camera may be turned off).

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

Part 1: Solve puzzles individually

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

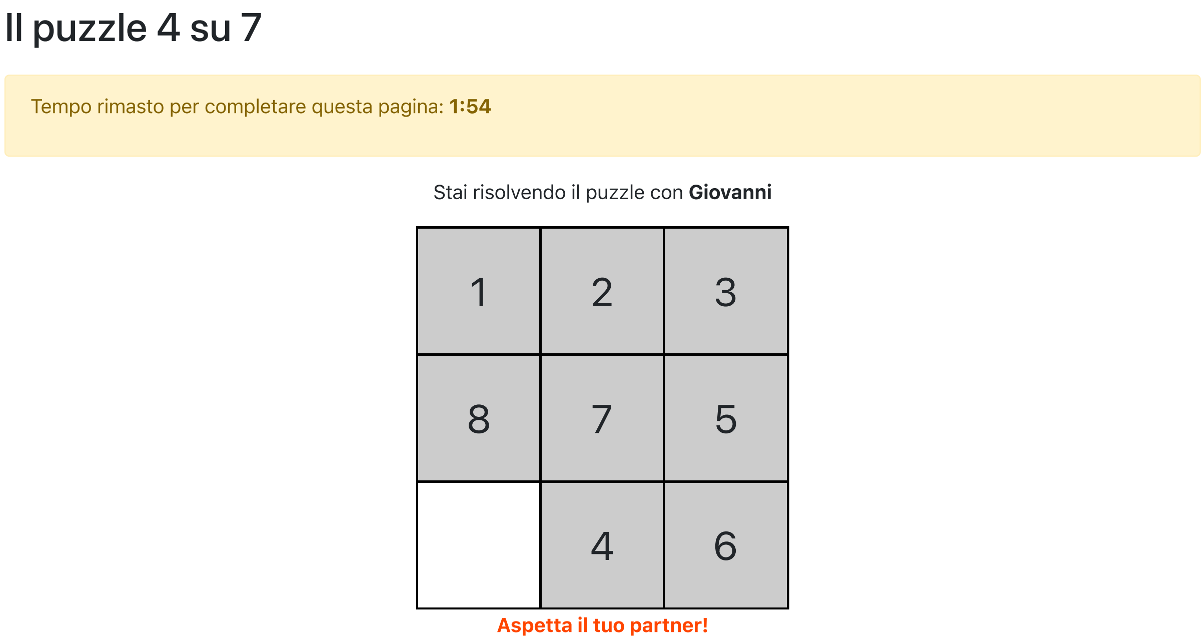
9. I draw with replacement a number from 1 to 40 using Google's random number generator (which is displayed by searching with "random number generator"). If no participant has a coin with the drawn number, I draw next number until the number of participants is 16. I share my computer screen so that participants see the numbers are actually drawn randomly.

Part 2: Select a collaborator

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

Second, participants are paired with another randomly chosen participant in the same group and solve one puzzle together by alternating their moves. The participant who makes the first move is drawn at random and both participants know this first-mover selection criterion. If they cannot solve the puzzle within 2 minutes, they finish the puzzle without solving it. Participants are allowed to reverse the paired participant's move.[10] Each participant's performances in a given puzzle are measured as defined in Appendix B. Figure 3 shows a sample puzzle screen where a participant is paired with another participant called Giovanni and waiting for Giovanni to make his move.

Figure 3: Puzzle screen



Notes: This figure shows a sample puzzle screen where a participant is matched with another participant called Giovanni at the 4th round puzzle and waiting for Giovanni to make his move.

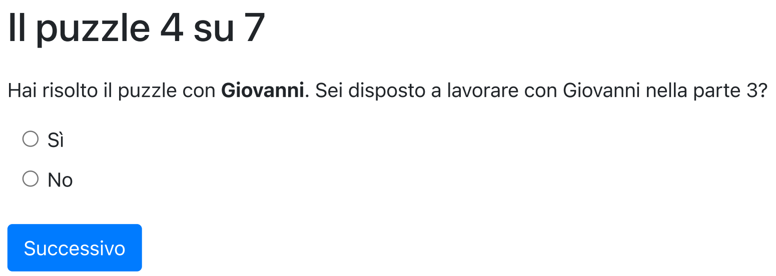
Once they finish the puzzle, participants state whether they would like to collaborate with the same participant in part 3 (yes/no). At the end of the first round, new pairs are formed, with a perfect stranger matching procedure, so that every participant is paired with each of the other 7 members of their group once and only once. In each round, participants solve another puzzle in a pair, then state whether they would like to collaborate with the same participant in part 3. The sequence of puzzles is the same for all pairs in all sessions. The puzzle difficulty is

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

kept the same across the seven rounds. The minimum number of moves to solve the puzzles is set to 8 based on the pilot.

The paired participant's first name is displayed on the computer screen throughout the puzzle and when participants select their collaborator to subtly inform the paired participant's gender. Figure 4 shows an example of the collaborator selection screen where a participant finished playing a puzzle with another participant called Giovanni and must state whether she or he would like to collaborate with Giovanni in part 3.

Figure 4: Collaborator selection screen



Notes: This figure shows a sample collaborator selection screen where a participant finished solving the 4th round puzzle with another participant called Giovanni and deciding whether she or he would like to collaborate with Giovanni in part 3.

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

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

Part 3: Solve puzzles with a collaborator

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

Post-experiment

Each participant answers a short questionnaire which consists of (i) the six hostile and benevolent sexism questions used in Stoddard, Karpowitz, and Preece (2020) with US college students and (ii) their basic demographic information and what they have thought about the experiment. The answer to the sexism questions is used to construct a gender bias measure (see Appendix C for the construction of the measure) and their demographic information is used to know participants' characteristics as well as casually check whether they have anticipated that the experiment is about gender.[11]

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

2.2 Implementation

The experiment was programmed with oTree (Chen, Schonger, and Wickens 2016) and conducted in Italian on a Heroku server and Zoom during November-December 2020. I recruited 464 participants (244 female and 220 male) registered on the Bologna Laboratory for Experiments in Social Science's ORSEE (Greiner 2015) who (i) were students, (ii) were born in Italy and (iii) had not participated in gender-related experiments before (as far as I could check).[12] The first two conditions were to reduce noise coming from differences in socio-demographic backgrounds and race or/and ethnicity that may be inferred from participants' first name or/and voice and the last condition was to reduce experimenter demand effects. The number of participants was determined by a power simulation in the pre-analysis plan to achieve 80% power.[13] The experiment is pre-registered with the OSF.[14]

I ran 29 sessions with 16 participants each. The average duration of a session was 70 minutes. The average total payment per participant was 11.55€ with the maximum 25€ and the minimum 2€, all including the 2€ show-up fee.

3 Data

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

3.1 Participants' characteristics

Table 1 describes participants' characteristics. Male participants are slightly older than female participants by 1.4 years and more gender-biased. People from southern Italy are slightly overrepresented for both female and male participants.[15] Female participants are more likely to major in humanities and male participants are more likely to major in natural sciences and engineering, a tendency observed in most OECD countries (see, for example, Carrell, Page, and West (2010)). Most female and male participants are either bachelor or master students (97% of female and 94% of male).

11. None has anticipated that the puzzle is about gender.
12. The laboratory prohibits deception, so no participant has participated in an experiment with deception.
13. This number includes 16 participants from a pilot session run before the pre-registration where the experimental instructions were slightly different. The results are robust to exclusion of these 16 participants.
14. The pre-registration documents are available at the OSF registry: https://osf.io/tgyc5.
15. Despite that I recruited only Italy-born people, 1 male participant answered in the post-questionnaire that he was from abroad. I include this participant in the analysis anyway but the results are robust to excluding this participant from the data.

3.2 Move-level summary

Figure 5 shows the average move quality along with 95% confidence intervals (panel A), the fraction of total moves in each move (panel B), and the probability of corrections in each move (panel C), separately for female only (gray), male only (white), and mixed gender pairs (blue).

Panel A shows that for all kinds of pairs, the average move quality is around 0.8 (8 out of 10 are good moves) until the 8th move (the minimum number of moves to solve a puzzle). After the 8th move, move quality deteriorates and stays around 0.6 (6 out of 10 are good moves). Panel B shows that for all kinds of pairs, about 71% of the puzzles are solved within 8 moves ((0.0875-0.025)/0.0875≈0.71), which is the minimum number of moves to solve the puzzle, then the other 30% takes more. Panel C shows that corrections happen across the moves, but are more likely to happen after the 8th move.

3.3 Puzzle-level summary

Table 2 describes own (panel A) and partner's puzzle-solving ability (panel B), corrections received (panel C), and puzzle outcomes (panel D). Panel A shows that female participants solve 0.44 fewer puzzles in part 1. However, there are no gender differences in contribution to the puzzle in part 2. This is likely because I grouped participants with similar abilities. Panel B shows that partner's puzzle-solving ability is not different when they are paired with female or male participants.[16]

Figure 6 shows the distribution of ability measures to elaborate panel A of Table 2. First, panel A shows that most participants contributed the same degree, in about 70% of the puzzles participants' contribution is 4 (total good moves minus total bad moves). Second, male participants seem to have solved more puzzles even in distribution, which may be reflected in their puzzle moves in part 2 not captured by contribution, for example, speed of making a move.

Panel C shows that participants are corrected by their partner in 15-16% of the total puzzles, of which 10-11% are corrections of mistakes and 5% are corrections of a right move.[17]

Panel D shows that participants state they want to collaborate with the partner 71-72% of the time. Participants spend on average 43-44 seconds for each puzzle (the maximum time a pair can spend is 120 seconds) and take 11 moves (remember the minimum number of moves to solve the puzzle is 8). 85-86% of the puzzles are solved and participants and the partner correct each other's move consecutively in 4% of the puzzles. There is no gender difference in any of these outcomes.

16. This definition of contribution is what Isaksson (2018) defines as "performance." In the pre-analysis plan, I defined i's contribution as "performance" of i divided by sum of "performance" of i and j and truncated values outside [0,1]. However, in my data, there is truncation in more than 10% of the puzzle, and the original contribution measure may not appropriately reflect participants' actual contribution. Thus, I instead use this "performance" measure in my analysis; since I add individual fixed effects, whether a measure is relative or absolute does not matter. However, the same results hold when I use original contribution measure, except table 6 where women's reaction to being corrected a mistake and a right move is statistically significant and men's reaction quantitatively and statistically more significant.

17. Of the 3180 puzzle, there are 495 puzzles where at least one correction occurred, of which 325 puzzles experienced only good corrections and 110 only bad corrections. The remaining 60 puzzles experienced both good and bad corrections. In order for good and bad corrections to capture only good and bad correction effect, I classify these 60 puzzles to good corrections if there were more good corrections than bad corrections (19 puzzles) and to bad corrections otherwise (41 puzzles). This classification is a bit arbitrary, but the results are robust to excluding these 60 puzzles, which I show in section 9.

3.4 Balance across rounds

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

First, panel A shows that partner gender is roughly balanced across rounds, except in the first round where female participants are less likely to face female partners and male participants more likely to female participants. Second, panels B-H show that most outcome variables are unbalanced across rounds both for female and male participants; specifically, whether a participant is selected as a collaborator and a puzzle is solved are lower in rounds 6 and 7. Also, while the number of corrections, time a pair spends on the puzzle, and total moves – all of which are likely to affect collaborator selection – are higher in rounds 6 and 7. It is unclear why there are these imbalances across rounds because all puzzles are the same difficulty: it could be that participants got tired in later rounds, puzzles in rounds 6 and 7 are perceived more difficult, etc.

However, they are all outcomes of a particular pair so they are just correlations. I show in section 9 that the results are robust to exclusion of rounds 6 and 7.

4 Theoretical framework

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

I consider a participant i who maximizes her or his expected utility by selecting their collaborator j from a set of i's potential collaborators $J \equiv \{1, 2, 3, 4, 5, 6, 7\}$. i's utility depends on her or his payoff and emotion. The utility is increasing in the payoff and the payoff is increasing in i's belief about j's ability. Thus, if i would select with whom to play in part 3, she or he would face the following problem:

$$\max_{j \in J} E_{\mu_j}[u_i(\underbrace{\pi(\mu_j(\tilde{a}_j, c_j))}_{\text{i's payoff}}, \underbrace{\kappa_i(c_j)}_{\text{i's emotion}})|\theta_i], \quad \partial u_i/\partial\pi > 0,\ \partial\pi/\partial\mu_j > 0 \tag{1}$$

where each term is defined as follows:

- μ_j: i's belief about j's ability
- $\tilde{a}_j$: j's ability perceived by i
- c_j: j's correction (=1 if j corrected i, =0 not corrected)
- θ_i: i's belief about her or his ability relative to other participants (>0 if high, =0 if same, <0 if low)

I assume:

- μ_j is increasing in j's ability perceived by i: $\partial\mu_j/\partial\tilde{a}_j > 0$
- i's utility is decreasing in her or his emotion: $\partial u_i/\partial\kappa_i < 0$
- emotion is irrelevant if i is fully rational: $u_i(\pi, \kappa_i) \propto u_i(\pi)$

If i can fully observe j's move quality and i is fully rational, then j's correction, c_j, does not convey any information about j's ability and is irrelevant for i's decision making. However, since i can only partially observe j's move quality, j's correction conveys information about j's ability even if i is fully rational.[18]

4.1 When i is fully rational

First, *keeping j's ability perceived by i ($\tilde{a}_j$) fixed*, as I do in the analysis, the information j's correction conveys depends on θ_i. If i believes she or he is good at the puzzle, she or he would consider a correction as a signal of low ability because i believes her or his move is correct. On the other hand, if i believes her or his ability is low, then she or he would consider a correction as a signal of high ability. If i believes his ability is the same as j's, then a correction would not convey any information. Thus,

- $\partial\mu_j/\partial c_j < 0$ if $\theta_i > 0$,
- $\partial\mu_j/\partial c_j = 0$ if $\theta_i = 0$, and
- $\partial\mu_j/\partial c_j > 0$ if $\theta_i < 0$.

4.2 When i is not fully rational

When i is not fully rational, i's emotion, κ_i, matters for her or his maximization problem. Specifically, I assume that j's correction induces i's negative feeling towards j: $\partial\kappa_i/\partial c_j < 0$.

The assumption is based on the literature on motivated reasoning (Kunda 1990). The first assumption is based on the finding that people consider those who disagree with them as biased (Kennedy and Pronin 2008). While it is a belief, I assume it affects i's actions.

5 Response to being corrected

In this section, I document evidence that people – both women and men – are less willing to work with a person who corrected their move after controlling for that person's contribution to the puzzle.

5.1 Response to being corrected: Estimating equation

I run the following OLS regression.

$$Select_{ij} = \beta_1 Corrected_{ij} + \beta_2 Female_j + \delta_1 Contribution_j + \delta_2 \#PuzzlesPt1_j + \mu_i + \epsilon_{ij} \quad (2)$$

where each variable is defined as follows:

- $Select_{ij} \in \{0, 1\}$: an indicator variable equals 1 if i selects j as their collaborator, 0 otherwise.
- $Corrected_{ij} \in \{0, 1\}$: an indicator variable equals 1 if i is corrected by j, 0 otherwise.
- $Female_j \in \{0, 1\}$: an indicator variable equals 1 if j is female, 0 otherwise.
- $Contribution_j \in \mathbb{Z}$: j's contribution to a puzzle played with i.
- $\#PuzzlesPt1_j \in \{0, 1, ..., 15\}$: number of puzzles j has solved in part 1.
- ϵ_{ij}: omitted factors that affect i's likelihood to select j as their collaborator.

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

and $\mu_i \equiv \sum_{k=1}^{N} \mu^k \mathbb{1}[i = k]$ is individual fixed effects, where N is the total number of participants in the sample and $\mathbb{1}$ is the indicator variable. Standard errors are clustered at the individual level.[19]

The key identification assumption is that $Contribution_j$ and $\#PuzzlesPt1_j$ fully capture j's ability *perceived* by i (not true ability).[20]

5.2 Response to being corrected: Results

Table 3 presents the regression results of equation 2. Columns 1, 3, and 5 show that when we do not control for partner's ability measures, the correction effect is significantly downward-biased, suggesting that the main determinant of participant's collaborator selection is the partner's contribution to the puzzle.

Looking at columns 2, 4, 6, and 7, the coefficient estimate on the partner's contribution is positive and highly significant both quantitatively and statistically. The coefficient estimate on the partner's number of puzzles solved in part 1 is also statistically significant, but only for female participants.

The coefficient estimate on being corrected in column 2 is negative and statistically significant, suggesting that people are less willing to work with a person who corrected their moves. This effect is present for both women (column 4) and men (column 6), but stronger for women (column 7).

6 Response to being corrected a mistake vs. a right move

In this section, I separate corrections of mistakes and right moves and document evidence that while women are less willing to work with a person who corrected their mistakes as well as right moves, men are mostly less willing to work with a person who corrected their mistakes.

6.1 A mistake vs. a right move: Estimating equation

I run the following OLS regression.

$$\begin{aligned} Select_{ij} = &\beta_1 CorrectedMistake_{ij} + \beta_2 CorrectedRightMove_{ij} + \beta_3 Female_j + \\ &\delta_1 Contribution_j + \delta_2 \#PuzzlesPt1_j + \mu_i + \epsilon_{ij} \end{aligned} \tag{3}$$

where each variable is defined as follows:

- $CorrectedMistake_{ij} \in \{0, 1\}$: an indicator variable equals 1 if i is corrected by j for their mistakes (a move that makes the puzzle further away from the solution), 0 otherwise.

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

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

- $CorrectedRightMove_{ij} \in \{0,1\}$: an indicator variable equals 1 if i is corrected by j for their right moves (a move that makes the puzzle closer to the solution), 0 otherwise.

other variables are as defined in equation 2.

6.2 A mistake vs. a right move: Results

Table 4 presents the regression results of equation 3. Looking at columns 1, 3, and 5 show that when we do not control for partner's ability measures, the effect of correction of a right move is severely downward biased. This is additional evidence that the main determinant of collaborator selection is the partner's contribution to the puzzle.

The coefficient estimate on being corrected for a mistake in column 2 is negative and statistically and economically highly significant, suggesting that people are less willing to work with a person who corrected their mistakes. This effect is present for both women (column 4) and men (column 6) with a similar magnitude (column 7).

The coefficient estimate on being corrected for a right move in column 2 is also negative and statistically and economically highly significant, suggesting that people are less willing to work also with a person who corrected their right moves. However, there is a gender difference in the magnitude of this effect. While the effect is the same magnitude as the effect of being corrected for mistakes for women (column 3), the effect is weaker – or even close to zero – than the effect of being corrected for mistakes for men (column 5). Comparing women and men, men respond to being corrected a right move less strongly than women (column 7).

7 Is the negative response to being corrected rational or emotional?

So far, I document evidence that both women and men are less willing to work with a person who corrected their moves, but while women respond equally negatively to corrections of their mistakes and their right moves, men respond less negatively to corrections of their right moves than of their mistakes. However, since the quality of corrections is not fully observable, it is unclear whether these negative responses are consistent with Bayesian updating (they consider the correction as a signal of a person's low ability) or due to their emotional irritation.

In this section, I document evidence that men's negative response to being corrected is due to their emotional irritation.

7.1 Rational or emotional? Estimating equation

I run the following OLS regression.

$$\begin{aligned} Select_{ij} = &\beta_1 CorrectedMistake_{ij} + \beta_2 CorrectedRightMove_{ij} + \\ &\beta_3 CorrectedMistake_{ij} \times HighAbility_i + \beta_4 CorrectedRightMove_{ij} \times HighAbility_i + \\ &\beta_5 Female_j + \delta_1 Contribution_j + \delta_2 \#PuzzlesPt1_j + \mu_i + \epsilon_{ij} \end{aligned} \tag{4}$$

where each variable is defined as follows:

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

other variables are as defined in equations 2 and 3.

7.2 Rational or emotional? Results

Figure 8 shows point estimates and 95% confidence intervals of being corrected for a mistake and being corrected for a right move for high-ability (blue) and low-ability participants (green) from the regression results of equation 4. Panel A shows that high-ability women are neither less nor more reluctant to being corrected for their mistakes or their right moves. However, panel B shows that high-ability men are more reluctant to being corrected for their mistakes while they are not less reluctant to being corrected for their right moves than low-ability men.

If the negative response to being corrected is due to their Bayesian updating, high-ability people should be able to observe move quality better than low-ability people and respond less negatively to corrections of their mistakes than to corrections of their right moves. Thus, the results suggest that for men, the negative response to being corrected is due to their emotional irritation.

8 Does the gender of the partner matter in response to being corrected?

A study finds that men view women in a (more) stereotypical way when women criticize them (Sinclair and Kunda 2000). Another study finds that people punish out-group members' misbehavior more than that of in-group members (Chen and Li 2009). There is also evidence that women are more harshly punished for their mistakes than men (Egan, Matvos, and Seru 2019; Sarsons 2019). Thus, it may be that men respond more negatively to women's corrections than to men's corrections.

In this section, I document that men do not respond more negatively to women's correction than to men's correction.

8.1 Does the gender of the partner matter? Estimating equation

I run the following OLS regression.

$$\begin{aligned} Select_{ij} = &\beta_1 Corrected_{ij} + \beta_2 Female_j + \beta_3 Corrected_{ij} \times Female_j \\ &+ \delta_1 Contribution_j + \delta_2 \#PuzzlesPt1_j + \mu_i + \epsilon_{ij} \end{aligned} \tag{5}$$

where each variable is defined as in equation 2.

8.2 Does the gender of the partner matter? Results

Figure 9 shows point estimates and 95% confidence intervals of being corrected, being corrected for a mistake, and being corrected for a right move separately for female (blue) and male partners (green) from the regression results of equation 5. Panel A shows that women do not respond more negatively to women's corrections than to men's corrections. Panel B also shows that men do not respond more negatively to women's corrections either; men's response to women's correction of a mistake seems more negative than that of men's, but the difference is not statistically significant.

9 External validity and robustness

In this section, I argue that the findings so far are likely to be lower bound and are robust to alternative explanations.

9.1 External validity

While the laboratory setting is different from the real-world workplace, my findings are likely to be lower bound because of the two reasons. First, being corrected is not observed by others in my experiment: those who have been corrected do not lose face in front of other people, unlike in the real-world workplace. Second, the emotional stake is much smaller: it is just a puzzle after all and not something people have been devoting much of their time to, such as research projects and corporate investment projects.

9.2 Robustness

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

To address this concern, I re-estimate equations 3 and 4 with solved puzzles only. Columns 1, 2, 5, and 6 of Table 5 present the results, which show that the findings are robust to excluding unsolved puzzles.

In addition, excluding unsolved puzzles makes the distribution of contribution tighter – panel A of figure 10 shows that in about 80% of the puzzles participants contributed the same degree. This makes it more credible that contribution appropriately captures paired partner's ability observed by the participants.

Excluding rounds 6 and 7 We see in figure 7 that participants are less willing to collaborate with the paired participants in rounds 6 and 7. Also, there are more corrections in rounds 6 and 7 than in other rounds. Although they are both outcomes of particular pairs, one may wonder whether rounds 6 and 7 are driving the results.

To address this concern, I re-estimate equations 3 and 4 with solved rounds 1-5 only. Columns 3, 4, 7, and 8 of Table 5 present the results, which show that the findings are robust to excluding rounds 6 and 7.

Excluding puzzles where both good and bad corrections occurred As discussed in the footnote of section 3.3, there are 495 puzzles in which at least one correction occurred, of which

325 puzzles experienced good corrections only, 110 puzzles bad corrections only, and 60 puzzles experienced both good and bad corrections. In puzzles that experienced both good and bad corrections, I considered that the puzzles experienced good corrections if there were more good corrections than bad corrections, and experienced bad corrections otherwise. However, some people may think classification is a bit arbitrary.

To address this concern, I re-estimate equations 3 and 4 with puzzles in which only good or bad corrections occurred. Table 6 presents the results, which show that the findings are robust to excluding puzzles where both good and bad corrections occurred.

10 Conclusions

This paper studies how being corrected by others in a group affects one's willingness to collaborate with that person in later works. I design a quasi-laboratory experiment where participants are paired with seven other participants, solve one number-sliding puzzle together, and express a preference on which of them to be paired with in the final, payoff-relevant, part of the experiment. I find that the paired participants' contribution to the puzzle is the most important factor for participants in selecting their collaborator. However, once I control for the paired participants' contribution to the puzzle, participants are significantly less likely to select a paired participant who corrected their move. Women do not like being corrected for their mistakes as well as their right moves, while men mostly do not like being corrected for their mistakes. High-ability men especially do not like to be corrected for their mistakes, suggesting that the emotional irritation is driving their negative reactions. The gender of the paired participants who make corrections does not matter for negative response to being corrected.

These findings have three implications. First, people's reluctance to being corrected can be a source of organizational inefficiency. Second, men's reluctance to being corrected for their mistakes but not for their right actions could discourage workers to correct their male leaders. Third, evidence on gender differences in response to being corrected enriches our understandings of gender differences in corrections and strategic behaviors.

References

Andreoni, James, and Lise Vesterlund. 2001. “Which is the Fair Sex? Gender Differences in Altruism.” *The Quarterly Journal of Economics* 116 (1): 293–312.

Apesteguia, Jose, Ghazala Azmat, and Nagore Iriberri. 2011. “The Impact of Gender Composition on Team Performance and Decision Making: Evidence from the Field.” *Management Science* 58 (1): 78–93.

Arechar, Antonio A., Simon Gächter, and Lucas Molleman. 2018. “Conducting interactive experiments online.” *Experimental Economics* 21 (1): 99–131.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. 2009. “Social Connections and Incentives in the Workplace: Evidence From Personnel Data.” *Econometrica* 77 (4): 1047–1094.

Beaman, Lori, and Jeremy Magruder. 2012. “Who Gets the Job Referral? Evidence from a Social Networks Experiment.” *American Economic Review* 102 (7): 3574–3593.

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

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

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.

Chen, Yan, and Sherry Xin Li. 2009. “Group Identity and Social Preferences.” *American Economic Review* 99 (1): 431–457.

Coffman, Katherine B., Clio Bryant Flikkema, and Olga Shurchkov. 2021. *Gender Stereotypes in Deliberation and Team Decisions.* Working Paper.

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

Dehdari, Sirus H., Emma Heikensten, and Siri Isaksson. 2019. *What Goes Around (Sometimes) Comes Around: Gender Differences in Retaliation.* Working Paper.

Eckel, Catherine C., and Philip J. Grossman. 1998. “Are Women Less Selfish Than Men?: Evidence From Dictator Experiments.” *The Economic Journal* 108 (448): 726–735.

———. 2001. “Chivalry and solidarity in ultimatum games.” *Economic Inquiry* 39 (2): 171–188.

Egan, Mark, Gregor Matvos, and Amit Seru. 2019. *When Harry Fired Sally: The Double Standard in Punishing Misconduct.* Working Paper.

Ertac, Seda, and Mehmet Y. Gurdal. 2012. “Deciding to decide: Gender, leadership and risk-taking in groups.” *Journal of Economic Behavior & Organization* 83 (1): 24–30.

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.

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

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

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. 2020. *Overriding in teams: The role of beliefs, social image, and gender.* Working Paper.

Hjort, Jonas. 2014. "Ethnic Divisions and Production in Firms." *The Quarterly Journal of Economics* 129 (4): 1899–1946.

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

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

Kennedy, Kathleen A., and Emily Pronin. 2008. "When Disagreement Gets Ugly: Perceptions of Bias and the Escalation of Conflict." *Personality and Social Psychology Bulletin* 34 (6): 833–848.

Kunda, Ziva. 1990. "The case for motivated reasoning." *Psychological Bulletin* 108 (3): 480–498.

Li, Xuan. 2020. *The Costs of Workplace Favoritism: Evidence from Promotions in Chinese High Schools.* Working Paper.

MacLeod, W. Bentley. 2003. "Optimal Contracting with Subjective Evaluation." *American Economic Review* 93 (1): 216–240.

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

Manian, Shanthi, and Ketki Sheth. 2021. "Follow my Lead: Assertive Cheap Talk and the Gender Gap." *Management Science.*

Matsa, David A., and Amalia R. Miller. 2013. "A Female Style in Corporate Leadership? Evidence from Quotas." *American Economic Journal: Applied Economics* 5 (3): 136–169.

Prendergast, Canice. 1993. "A Theory of "Yes Men"." *American Economic Review* 83 (4): 757–770.

Prendergast, Canice, and Robert Topel. 1996. "Favoritism in Organizations." *Journal of Political Economy* 104 (5): 958–78.

Reifen Tagar, Michal. 2014. *Why Disagreement Obstructs Constructive Dialogue: The Role of Biased Attribution of Moral Motives.* PhD Dissertation. University of Minnesota.

Reuben, Ernesto, Pedro Rey-Biel, Paola Sapienza, and Luigi Zingales. 2012. "The emergence of male leadership in competitive environments." *Journal of Economic Behavior & Organization* 83 (1): 111–117.

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

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

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

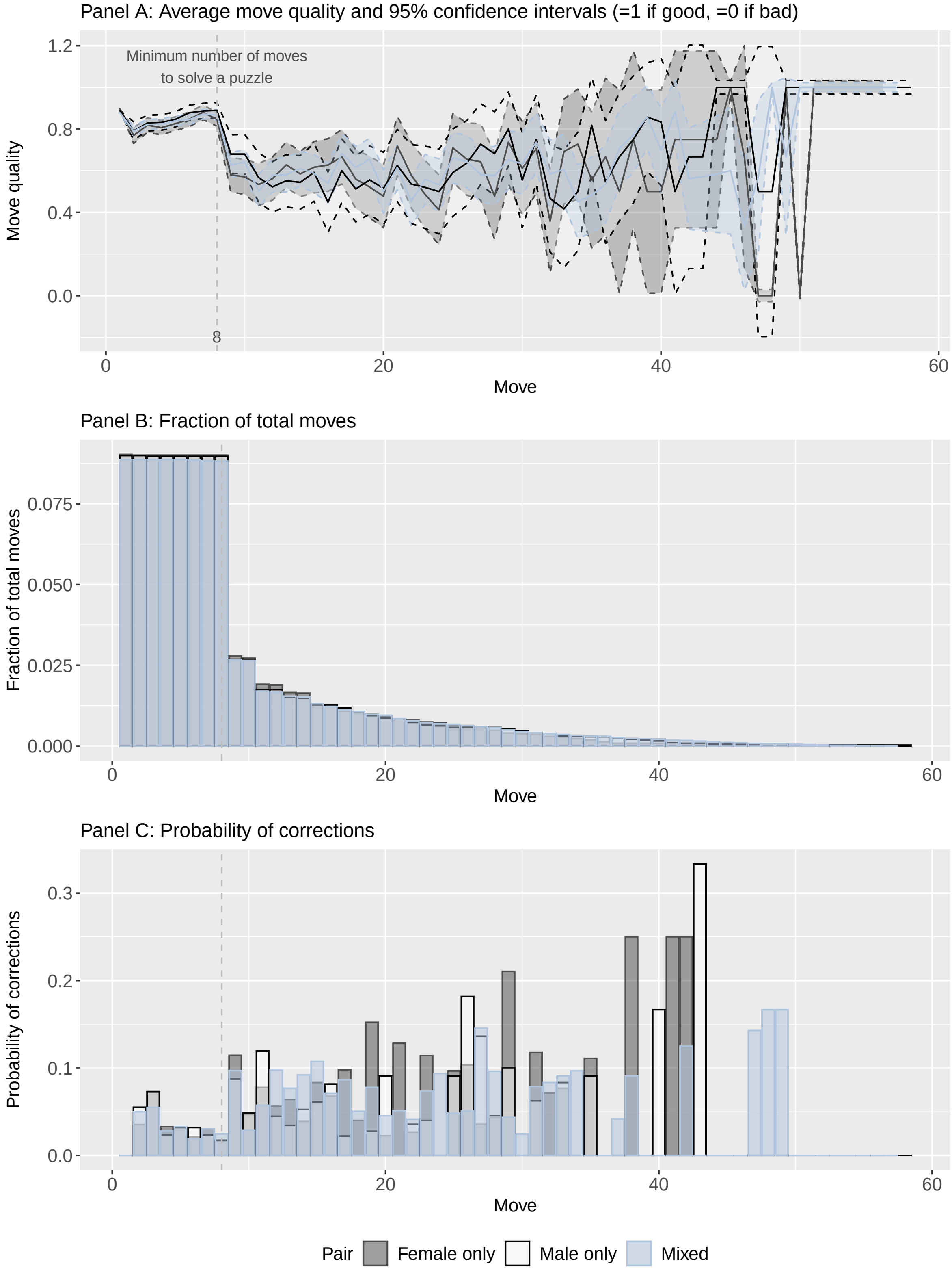
Solnick, Sara J. 2001. “Gender differences in the ultimatum game.” *Economic Inquiry* 39 (2): 189–200.

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

Xu, Guo. 2018. “The Costs of Patronage: Evidence from the British Empire.” *American Economic Review* 108 (11): 3170–3198.

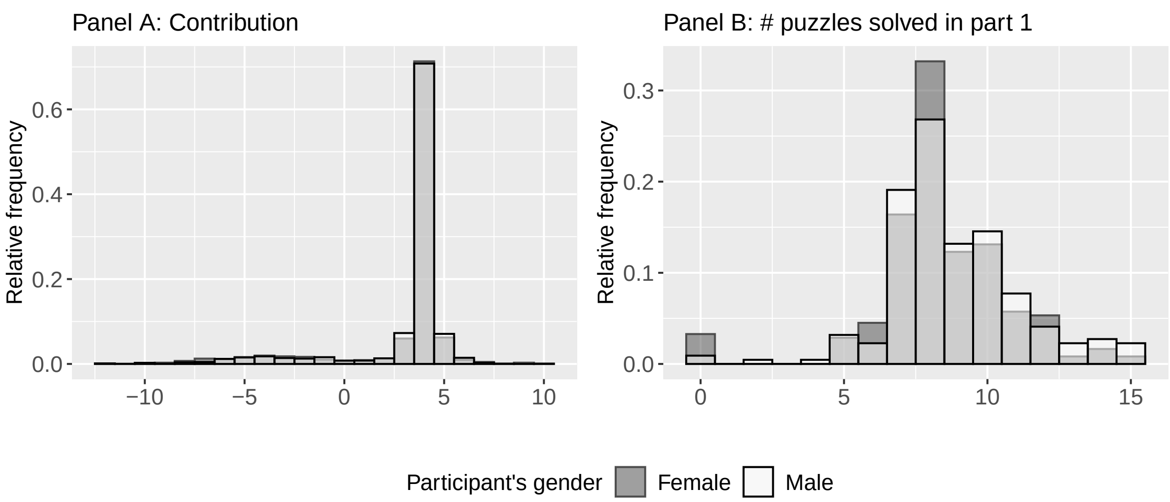
Figures

Figure 5: Move quality, fraction of total moves, and probability of corrections



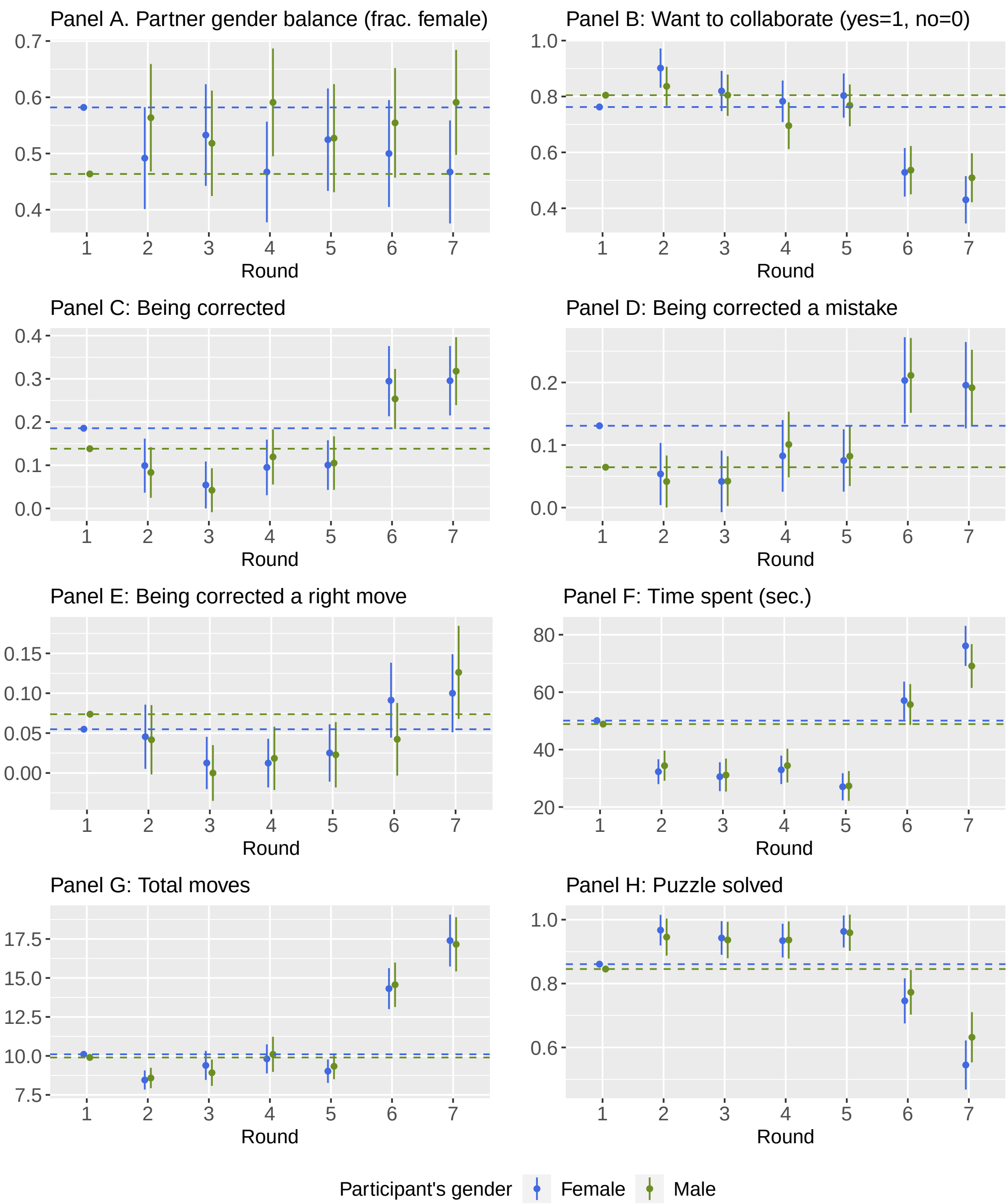
Notes: The average move quality along with 95% confidence intervals (panel A), the fraction of total moves in each move (panel B), and the probability of corrections in each move (panel C), separately for female only (gray), male only (white), and mixed gender pairs (blue). The confidence interval of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

Figure 6: Distribution of puzzle-solving ability



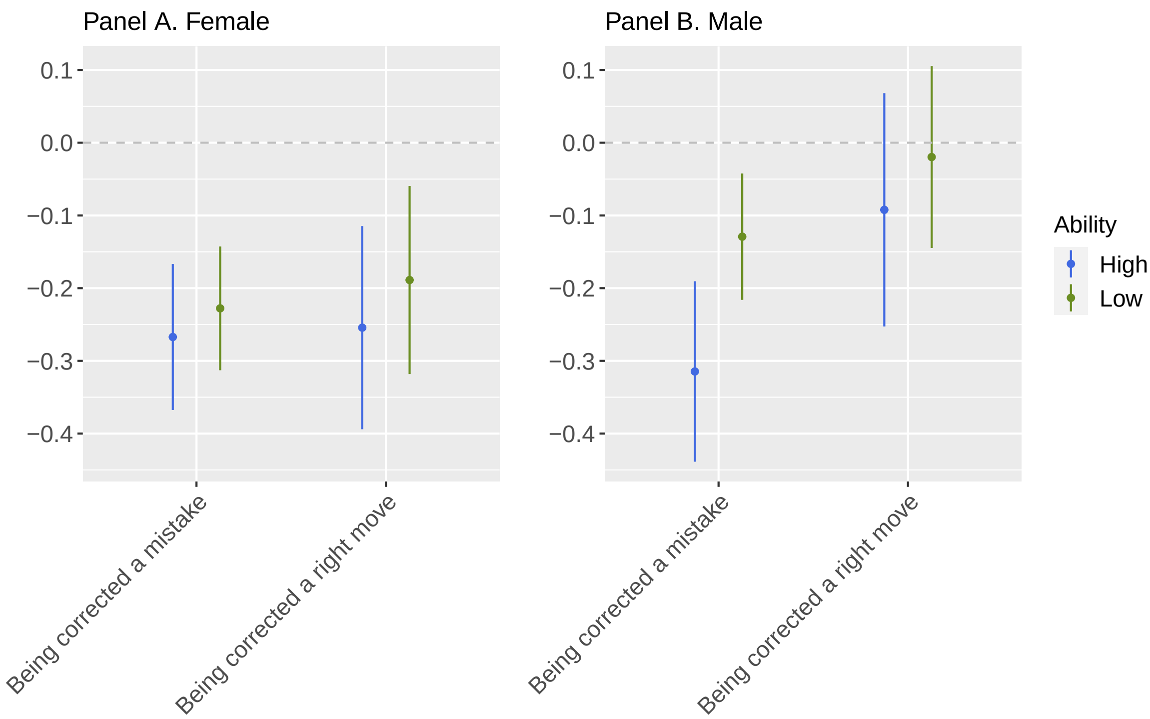
Notes: This figure shows the distribution of ability measures separately for female (gray) and male (white) participants. Appendix B provides definitions of each ability measure.

Figure 7: Balance across rounds



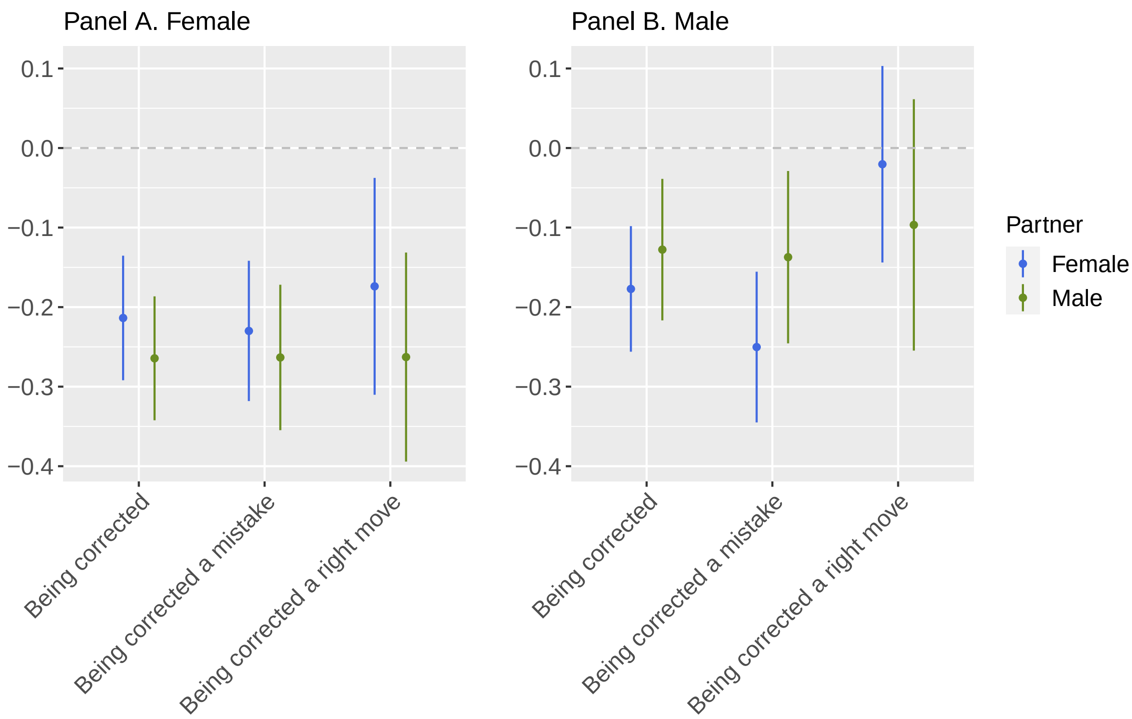
Notes: This figure shows point estimates and 95% confidence intervals of βs from the following OLS regression with gender balance (female dummy) and different puzzle outcomes separately for female (blue) and male participants (green): $y_{ij} = \beta_1 + \sum_{k=2}^{7} \beta_k \mathbb{1}[t_{ij} = k]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

Figure 8: Is negative response to being corrected rational or emotional?



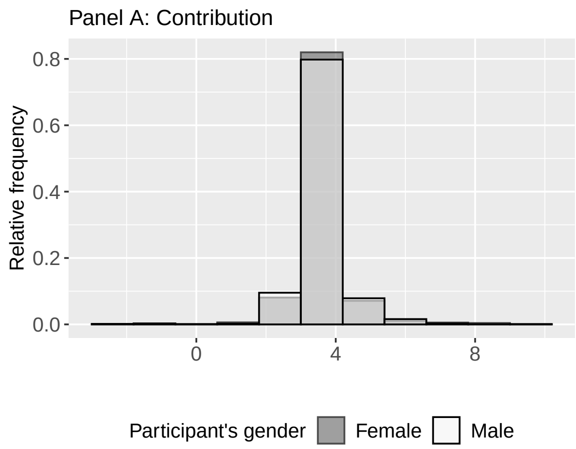
Notes: This figure shows point estimates and 95% confidence intervals of being corrected for a mistake and being corrected for a right move separately for for high-ability (blue) and low-ability participants (green) from the regression results of equation 4. It shows that men's negative response to being corrected is due to their emotional irritation. Standard errors are clustered at the individual level.

Figure 9: Does the gender of the partner matter in response to being corrected?



Notes: This figure shows point estimates and 95% confidence intervals of being corrected, being corrected for a mistake, and being corrected for a right move separately for female (blue) and male partners (green) from the regression results of equation 5. It shows that that men or women do not respond more negatively to women's correction than to men's correction. Standard errors are clustered at the individual level.

Figure 10: Excluding unsolved puzzles makes contribution less variable



Notes: This figure shows the distribution of contribution separately for female (gray) and male (white) participants, excluding unsolved puzzles, and shows that excluding unsolved puzzles makes contribution less variable and makes it more credible that contribution appropriately captures paired partner's ability observed by the participants.

Tables

Table 1: Participants' characteristics

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

Notes: This table describes participants' characteristics. Gender bias is measured with the 6 hostile and benevolent sexism questions and constructed as in Appendix C. P-values of the difference between female and male participants are calculated with heteroskedasticity-robust standard errors.

Table 2: Puzzle-solving ability, corrections, and puzzle outcomes

	Female (N=1708)		Male (N=1540)		Difference (Female – Male)		
	Mean	SD	Mean	SD	Mean	SE	P-value
Panel A: Own ability							
Contribution	2.98	2.93	3.14	2.64	-0.16	0.10	0.11
# puzzles solved in pt. 1	8.36	2.41	8.80	2.34	-0.44	0.22	0.05
Panel B: Partner's ability							
Contribution	3.04	2.73	3.07	2.87	-0.03	0.10	0.77
# puzzles solved in pt. 1	8.58	2.35	8.57	2.43	0.01	0.16	0.93
Panel C: Corrections							
Being corrected	0.16	0.37	0.15	0.36	0.01	0.01	0.51
Being corrected a mistake	0.11	0.31	0.10	0.31	0.01	0.01	0.59
Being corrected a right move	0.05	0.21	0.05	0.21	0.00	0.01	0.77
Panel D: Puzzle outcomes							
Want to collaborate (yes=1, no=0)	0.72	0.45	0.71	0.45	0.01	0.02	0.49
Time spent (sec.)	43.74	36.15	42.99	35.76	0.74	1.28	0.56
Total moves	11.18	7.46	11.21	7.70	-0.03	0.28	0.92
Puzzle solved	0.85	0.36	0.86	0.35	-0.01	0.01	0.43
Consecutive correction	0.04	0.20	0.04	0.21	0.00	0.01	0.81

Notes: This table describes own (panel A) and partner's puzzle-solving ability (panel B), corrections received (panel C), and puzzle outcomes (panel D). P-values of the difference between female and male participants are calculated with Standard errors clustered at the individual level. Appendix B provides definitions of each puzzle-solving ability measure.

Table 3: Response to being corrected

Outcome:	Want to collaborate (yes=1, no=0)						
Sample:	All		Female		Male		All
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Being corrected	-0.367***	-0.200***	-0.406***	-0.239***	-0.322***	-0.154***	-0.250***
	(0.025)	(0.022)	(0.037)	(0.030)	(0.033)	(0.031)	(0.029)
Female partner	-0.006	0.013	-0.013	0.006	0.003	0.019	0.012
	(0.017)	(0.014)	(0.022)	(0.018)	(0.026)	(0.022)	(0.014)
Partner's contribution		0.083***		0.089***		0.077***	0.083***
		(0.003)		(0.004)		(0.003)	(0.003)
Partner's # puzzles solved in pt. 1		0.010***		0.012**		0.007	0.010***
		(0.004)		(0.005)		(0.006)	(0.004)
Being corrected x Male							0.109***
							(0.041)
Individual FE	✓	✓	✓	✓	✓	✓	✓
Baseline mean	0.780	0.780	0.780	0.780	0.778	0.778	0.780
Baseline SD	0.414	0.414	0.414	0.414	0.416	0.416	0.414
Adj. R-squared	0.076	0.335	0.078	0.367	0.076	0.306	0.337
Observations	3180	3180	1670	1670	1510	1510	3180
Clusters	464	464	244	244	220	220	464

Notes: This table presents regression results of equation 2 and shows that both women and men are less willing to work with a person who corrected their moves, but women respond stronger to being corrected. Columns 1, 3, and 5 exclude partner's puzzle-solving ability to show that partner's contribution is the main determinant of partner selection: coefficient estimate on being corrected is significantly downward biased if we do not control for the ability. Baseline mean and standard deviation are that of partners who do not make corrections. Standard errors in parentheses are clustered at the individual level. Significance levels: * 10%, ** 5%, and *** 1%.

Table 4: Response to being corrected a mistake vs. a right move

Outcome:	Want to collaborate (yes=1, no=0)						
Sample:	All		Female		Male		All
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Being corrected a mistake	-0.267***	-0.224***	-0.304***	-0.247***	-0.223***	-0.195***	-0.250***
	(0.031)	(0.025)	(0.046)	(0.034)	(0.040)	(0.037)	(0.033)
Being corrected a right move	-0.580***	-0.138***	-0.634***	-0.217***	-0.522***	-0.049	-0.238***
	(0.036)	(0.038)	(0.048)	(0.051)	(0.052)	(0.053)	(0.047)
Female partner	-0.003	0.012	-0.011	0.006	0.006	0.018	0.012
	(0.017)	(0.014)	(0.022)	(0.018)	(0.026)	(0.022)	(0.014)
Partner's contribution		0.085***		0.090***		0.080***	0.085***
		(0.003)		(0.004)		(0.004)	(0.003)
Partner's # puzzles solved in pt. 1		0.010***		0.012**		0.007	0.010**
		(0.004)		(0.005)		(0.006)	(0.004)
Being corrected a mistake x Male							0.055
							(0.050)
Being corrected a right move x Male							0.217***
							(0.065)
Individual FE	✓	✓	✓	✓	✓	✓	✓
Being corrected a mistake	0.313***	-0.086**	0.330***	-0.030	0.299***	-0.146**	
−Being corrected a right move	(0.047)	(0.042)	(0.065)	(0.056)	(0.066)	(0.063)	
Baseline mean	0.780	0.780	0.780	0.780	0.778	0.778	0.780
Baseline SD	0.414	0.414	0.414	0.414	0.416	0.416	0.414
Adj. R-squared	0.092	0.336	0.096	0.367	0.089	0.309	0.339
Observations	3180	3180	1670	1670	1510	1510	3180
Clusters	464	464	244	244	220	220	464

Notes: This table presents regression results of equation 3 and shows that while women are less willing to work with a person who corrected their mistakes as well as right moves, men are mostly less willing to work with a person who corrected their mistakes. Baseline mean and standard deviation are that of partners who do not make corrections. Standard errors in parentheses are clustered at the individual level. Significance levels: * 10%, ** 5%, and *** 1%.

Table 5: Results are robust to exclusion of unsolved puzzles and rounds 6 and 7

Outcome:	Want to collaborate (yes=1, no=0)							
Sample:	Female, Solved puzzles		Female, Rounds 1-5		Male, Solved puzzles		Male, Rounds 1-5	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Being corrected a mistake	-0.311***	-0.285***	-0.189***	-0.127**	-0.369***	-0.275***	-0.225***	-0.099
	(0.052)	(0.066)	(0.046)	(0.058)	(0.056)	(0.067)	(0.062)	(0.064)
Being corrected a right move	-0.220**	-0.144	-0.165**	-0.066	-0.034	-0.030	-0.115	-0.092
	(0.098)	(0.126)	(0.083)	(0.102)	(0.090)	(0.120)	(0.079)	(0.111)
Female partner	-0.001	-0.001	-0.001	-0.001	0.006	0.006	0.016	0.018
	(0.020)	(0.020)	(0.020)	(0.020)	(0.023)	(0.023)	(0.025)	(0.024)
Partner's contribution	0.161***	0.160***	0.105***	0.107***	0.190***	0.187***	0.084***	0.086***
	(0.031)	(0.031)	(0.006)	(0.006)	(0.034)	(0.033)	(0.006)	(0.006)
Partner's # puzzles solved in pt. 1	0.010*	0.010*	0.010*	0.010*	0.010	0.010*	0.013*	0.014*
	(0.006)	(0.006)	(0.005)	(0.005)	(0.006)	(0.006)	(0.007)	(0.007)
Being corrected a mistake x High ability		-0.045		-0.129		-0.217**		-0.397***
		(0.094)		(0.091)		(0.097)		(0.132)
Being corrected a right move x High ability		-0.157		-0.238		-0.012		-0.036
		(0.149)		(0.156)		(0.155)		(0.144)
Individual FE	✓	✓	✓	✓	✓	✓	✓	✓
Being corrected a mistake	-0.091		-0.024		-0.335***		-0.110	
−Being corrected a right move	(0.125)		(0.093)		(0.121)		(0.099)	
Baseline mean	0.776	0.776	0.778	0.778	0.772	0.772	0.783	0.783
Baseline SD	0.417	0.417	0.416	0.416	0.420	0.420	0.412	0.412
Adj. R-squared	0.107	0.107	0.264	0.267	0.131	0.135	0.209	0.219
Observations	1449	1449	1199	1199	1321	1321	1083	1083
Clusters	244	244	244	244	220	220	220	220

Notes: This table presents regression results of equations 3 and 4 and shows that the findings so far are robust to exclusion of unsolved puzzles and rounds 6 and 7. Baseline mean and standard deviation are that of partners who do not make corrections. Standard errors in parentheses are clustered at the individual level. Significance levels: * 10%, ** 5%, and *** 1%.

Table 6: Results are robust to exclusion of puzzles where both good and bad corrections occurred

Outcome:	Want to collaborate (yes=1, no=0)			
Sample:	Female, No good-bad overlap		Male, No good-bad overlap	
	(1)	(2)	(3)	(4)
Being corrected a mistake	-0.224***	-0.176***	-0.193***	-0.123***
	(0.034)	(0.042)	(0.038)	(0.044)
Being corrected a right move	-0.138**	-0.113	-0.068	-0.034
	(0.058)	(0.071)	(0.061)	(0.078)
Female partner	0.003	0.002	0.018	0.017
	(0.019)	(0.019)	(0.022)	(0.022)
Partner's contribution	0.094***	0.094***	0.082***	0.082***
	(0.004)	(0.004)	(0.004)	(0.004)
Partner's # puzzles solved in pt. 1	0.011**	0.011**	0.007	0.007
	(0.005)	(0.005)	(0.006)	(0.006)
Being corrected a mistake x High ability		-0.099		-0.194**
		(0.067)		(0.078)
Being corrected a right move x High ability		-0.058		-0.076
		(0.117)		(0.114)
Individual FE	✓	✓	✓	✓
Being corrected a mistake	-0.086		-0.125*	
−Being corrected a right move	(0.065)		(0.069)	
Baseline mean	0.780	0.780	0.777	0.777
Baseline SD	0.415	0.415	0.416	0.416
Adj. R-squared	0.345	0.346	0.292	0.295
Observations	1633	1633	1487	1487
Clusters	244	244	220	220

Notes: This table presents regression results of equations 3 and 4 and shows that excluding the puzzles where both good and bad corrections occurred does not alter the results. Baseline mean and standard deviation are that of partners who do not make corrections. Standard errors in parentheses are clustered at the individual level. Significance levels: * 10%, ** 5%, and *** 1%.

Appendix A Pros and cons of quasi-laboratory format

On top of logistical convenience and complying with the COVID pre-caution measures, the quasi-laboratory format has an additional benefit over physical laboratory experiments in that participants cannot see each other when they enter the laboratory which adds an additional layer of anonymity among participants. A drawback is that participants can potentially be distracted while participating.

However, unlike standard online experiments such as on MTurk and Prolific where participants' identity is fully anonymous by the platforms' rule, we have participants' personal information and participants know it as we recruit them from our standard laboratory subject pool. Also, they are connected to us via Zoom throughout the experiment. These mostly prevent participants' attrition that can be endogenous to their decisions or treatments and the main problem of online interactive experiments (Arechar, Gächter, and Molleman 2018) and experiments where treatments affect the probability of attrition, e.g., experiments with intertemporal decision making. In my experiment, we experienced no participant attrition. A drawback is that we could not collect a large number of observations.

Another benefit of quasi-laboratory experiments over standard online experiments is that we can screen participants based on their participation status in previous experiments. This allows us to collect cleaner data; in particular, this allows us to screen out participants who have participated in experiments with deception, which is another problem of online experiments (Arechar, Gächter, and Molleman 2018).

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

Appendix B Definition of performance measures

Contribution I define a participant's contribution as their net good moves in a given puzzle in part 2:

$$\text{Player i's contribution} \equiv \text{i's \# good moves} - \text{i's \# bad moves} \in \mathbb{Z} \tag{B1}$$

The number of puzzles solves alone The number of puzzles a participant solves in part 1 of the experiment. Thus, it takes an integer value between 0 to 15.

Appendix C Construction of the gender bias measure

I construct the gender bias measure following Stoddard, Karpowitz, and Preece (2020) who use the measure to measure sexism of US undergraduate students.

As discussed in section 2.1, I ask participants to answer the following six hostile and benevolent sexism questions Stoddard, Karpowitz, and Preece (2020) have chosen from Glick and Fiske (1996)'s full-length sexism questionnaire.

Instructions: 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.

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

Answer choices to each question: Strongly agree, Agree a little, Neither agree nor disagree, Disagree a little, Strongly disagree

I assign a value of 4 to "Strongly agree," 3 to "Agree a little," 2 to "Neither agree nor disagree," 1 to "Disagree a little," and 0 to "Strongly disagree." Then I sum up the values for each participant and divide the sum by 24 which is the highest value one can receive. Thus, the measure takes a value from 0 to 1, and the higher the measure, the more gender-biased the person is. In the experiment, I use a certified Italian translation from Manganelli Rattazzi, Volpato, and Canova (2008) and Rollero, Glick, and Tartaglia (2014).