Gender Differences in the Cost of Corrections in Group Work

Yuki Takahashi*

Preliminary draft. Please do not cite or circulate.

Click here for the latest version

May 17, 2021

Abstract

Women speak up less often than equally knowledgeable men in a group, which reduces women's visibility in a group and group efficiency. However, speaking up corrects others who have a different opinion, and women's corrections may receive backlash. Should women speak up more often to close the labor market gender gap? This paper studies women's cost of correcting male group members and its consequence to group efficiency in a setting where the quality of corrections is only partially observable, as in most group work. I design a quasi-laboratory experiment where participants first perform a joint task seven times, each time with a different participant. After performing a joint task, they state whether they would like to be paired again with each of them. Then, they play a final, payoff-relevant, round of the task with one of the participants they have previously selected. After controlling for paired participants' contribution to the joint task and showing that statistical discrimination against women is unlikely to be present, I find that participants are significantly less likely to select a paired participant who has corrected their action, regardless of the paired participant's gender. Moreover, male participants react more negatively to a correction that fixes their mistake due to their overconfidence. These findings suggest that corrections do not necessarily increase group efficiency due to behavioral bias, it may not be necessarily optimal for women to speak up more, and that men may be speaking up too much.

JEL codes: J16, D91, C92
Keywords: correction, speak up, gender gap, group work

*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, 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 WEAI, PhD-EVS, and seminars at Ca' Foscari University, Catholic University of Brasília, the NHH, and the University of Bologna. 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

Most workplaces involve group work (Lazear and Shaw 2007) and we need to speak up so that we can bring together our expertise and make a better decision as a group. Speaking up is also important for us because individuals are not necessarily visible in a group. Women speak up less often than equally knowledgeable men (Coffman 2014); for example, in academia, women ask fewer questions than men during seminars (Carter et al. 2018; Dupas et al. 2021). This situation is suboptimal for the group and women themselves and may contribute to the labor market gender gap.

However, when we speak up, we correct others who have different opinions. Anecdotal evidence suggests that people find it difficult to correct others, especially those in senior positions.[1] This hesitance seems quite natural because correcting others essentially means telling others that they are wrong (Isaksson 2018). This is especially a problem because many times we could not see whether a correction is right or wrong and that women's correction may receive backlash (Rudman and Phelan 2008). Thus, it is unclear whether women should speak up more. If people corrected by a woman may not only fail to accept her correction but may also consider her as a less pleasant colleague to work with, she will have a problem in her career success because being selected into a team is essential to produce good outcomes in group work. Even in academia where one's work is more individualistic, one can have more projects on her portfolio when her colleagues prefer her as a co-author which increases their chance to get tenured.

This paper studies women's cost of correcting male colleagues and its efficiency consequence to the group. There are three main empirical challenges to investigate these questions 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. Third, a relationship between a correction and its outcome relevant in the workplace setting is difficult to measure.

To address these challenges, 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 be paired again 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 partner as possible. Participants are paired with all the seven group members in a random order to address endogenous group formation.

I use Isaksson (2018)'s number-sliding puzzle as the joint task where participants solve a 3x3 number-sliding puzzle in pairs by alternating their moves. I define a correction as reversing a group member's move to make different corrections comparable. The puzzle also 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, which makes it possible to keep individual contribution fixed and to determine whether a given correction improves group efficiency or not. Further, the task captures an essential characteristic of group work: two or more people work together

1. For example, check the following career magazine articles: Ashford (n.d.), Boogaard (n.d.), Greenwald (2018), Lebowitz and Akhtar (2019), Marshall (2020), McCord (n.d.), and Rosenberg McKay (2019).

towards the same goal (Isaksson 2018) but the correctness of each move and correction is only partially observable (albeit fully observable to the researcher). The outcome measure is whether a participant is selected as a partner, which is highly relevant in real-world group work.

I find that the main determinant of group members' partner selection is paired participants' contribution to the puzzle, they are equally likely to select women and men as a partner in absence of corrections and there are no gender differences in the contribution, suggesting that the partner selection is well-incentivized, group members partially observe paired participants' ability through their moves, and that statistical discrimination is unlikely to be present. There is no gender difference in the propensity to correct group member's moves either. However, after controlling for the paired participants' contribution, both male and female group members are less likely to select a paired participant who corrected their move as a partner by 12.2 percentage points or 16.3% relative to the baseline mean. This is economically significant: one has to increase her or his contribution to the puzzle by 0.59 standard deviation to offset this negative reaction. Yet, group members react equally negatively to women's and men's corrections.

This reluctance to accept being corrected may not reduce group efficiency if it is only about bad corrections that reverse a good move. However, this is not the case; in fact, male group members react more negatively to a correction that corrects their bad move than to a correction that corrects their good move. Female group members also react more negatively to a good correction, but to a much lesser extent. Thus, while group members appreciate the contribution part of the correction, they, especially men, dislike the part that points out their mistakes, hence missing an opportunity to select a good partner and a successful collaboration opportunity.

So why are group members reluctant to accept being corrected? Since the quality of the correction is not fully observable, corrections convey information about the paired participants' puzzle-solving ability. Thus, it is possible that group members believe that their move is correct and consider a correction of their move as a signal of the paired participant's low ability; in other words, it is group members' overconfidence that makes them reluctant to accept being corrected. This is indeed the case: group members who solved more puzzles in the individual practice stage (which precedes the partner selection stage) react more negatively to corrections than group members who solved fewer puzzles, keeping paired participants' ability fixed. However, this is not explained by their ability difference: group members who solved more puzzles in the individual practice stage respond more negatively to both good and bad corrections.

Taken together, these findings suggest that behavioral bias prevents groups to achieve efficiency gain from corrections. Because of that, speaking up more in a group can have a potential drawback for women: although women's corrections do not receive backlash, the negative effect of correction is pretty sizable. Also, men may be speaking up too much which is reducing their and group's efficiency.

Related literature This paper primarily relates to studies on gender differences in the contribution of ideas in group work. Coffman (2014) finds that women are less likely to contribute their ideas to the group in a male task due to self-stereotyping and Gallus and Heikensten (2019) find that debiasing their self-stereotyping by giving an award for their high ability increases women's contribution of their ideas: they put women's idea further ahead without involving open correction of their group member. However, on some occasions, the contribution of ideas has to be made openly, for example in academic seminars and business meetings. In such cases, group members' response plays an important role in the effectiveness of the intervention. Coffman, Flikkema, and Shurchkov (2021) find that group members are less likely to choose women's answers as a group answer in male-typed questions. Guo and Recalde (2020) find that group members correct women's ideas more often than men's ideas. Dupas et al. (2021) find that female economists receive more patronizing and hostile questions during seminars. Isaksson (2018) finds that men are more likely to correct their group member's bad moves in the same puzzle used in my experiment (as the puzzle was originally used by Isaksson (2018)). My paper introduces correction in the contribution of ideas and examines its cost and its effect on group efficiency.

More generally, my paper contributes to the literature on gender differences in group work. Isaksson (2018) finds that women under-claim their contribution compared to men in group work despite their equal contribution. Haynes and Heilman (2013) find similar results. Sarsons et al. (2021) find that people attribute less credit to a female economist when she co-authors a paper with a male economist(s). Born, Ranehill, and Sandberg (2020) and Stoddard, Karpowitz, and Preece (2020) find that women are less willing to lead a male-majority group. Shan (2020) finds that female students are more likely to drop out from an introductory economics class when they are assigned to a male-majority study group. Babcock et al. (2017) find that women are more likely to volunteer and be asked to do non-promotable tasks. My paper promotes our understanding of gender differences in group work.

My paper also speaks to the literature on social incentives in an organization (Ashraf and Bandiera 2018), in particular managerial favoritism. Literature develops theories that workers tend to conform their managers (Prendergast 1993) and managers favor workers whom they like in compensation and promotion (MacLeod 2003; Prendergast and Topel 1996) when objective worker performance measures are not available, both of which distorts the optimal allocation of talent. Several empirical studies verify these theoretical predictions (Bandiera, Barankay, and Rasul 2009; Beaman and Magruder 2012; Hjort 2014; Xu 2018). In addition, Li (2020) finds that managers' favoritism not only distorts the optimal allocation of talent but also reduces non-favored workers' performance. My paper suggests that people's reluctance to accept being corrected can lead to another form of distortion.

2 Experiment

As discussed in the introduction, there are three main challenges to study my research questions using secondary data: (i) endogenous group formation, (ii) difficulty to make different corrections comparable, and (iii) difficulty in measuring a relationship between a correction and its outcome relevant in the workplace setting. Thus, I answer my research questions in a controlled quasi-laboratory experimental setting.

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 own computers.[2] On top of logistical convenience and complying with the COVID precaution measures, this quasi-laboratory format has an additional benefit over physical laboratory experiments 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 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).

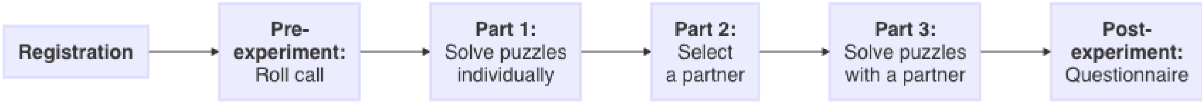
Group task As the group task I use Isaksson (2018)'s puzzle, a sliding puzzle with 8 numbered tiles, which should be placed in numerical order within a 3x3 frame (see figure 3 for an example). To achieve this goal, participants play in pairs, alternating their moves. This puzzle has nice mathematical 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.[3] 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.

2. There are already a few other studies that use a quasi-laboratory format, for example, Goeschl, Oestreich, and Soldà (2021). Michela Boldrini and Boon Han Koh have also conducted their experiments with a quasi-laboratory format, although their working papers are not yet publicly available.

3. 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.

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

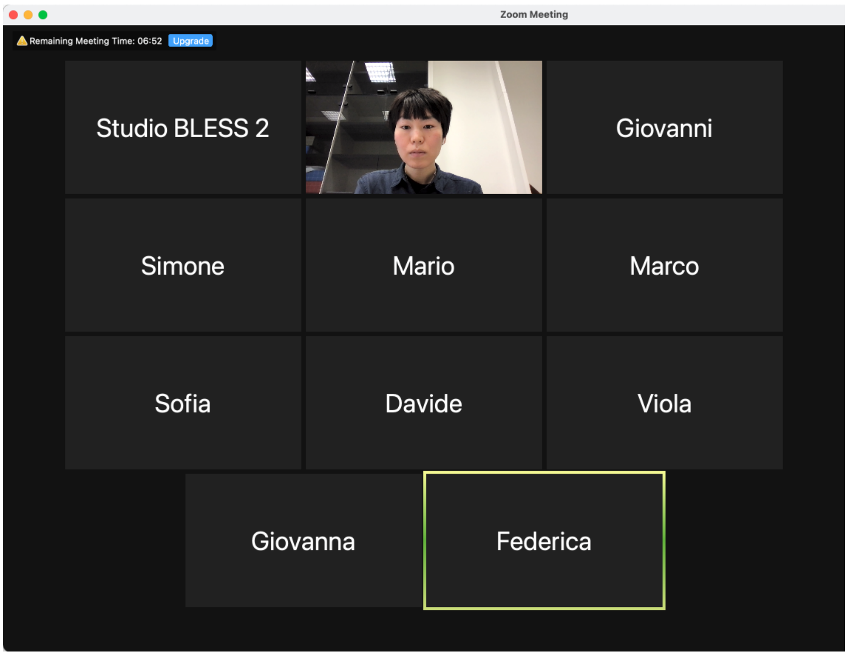
Upon receiving an invitation email to the experiment, participants register for a session they want to participate in and upload their ID documents as well as a signed consent form.[4]

Pre-experiment

On the day and the time of the session they have registered for, participants enter the Zoom waiting room.[5] 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).

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.

4. 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.

5. 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.

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 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.[6] 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.

Part 2: Select a partner

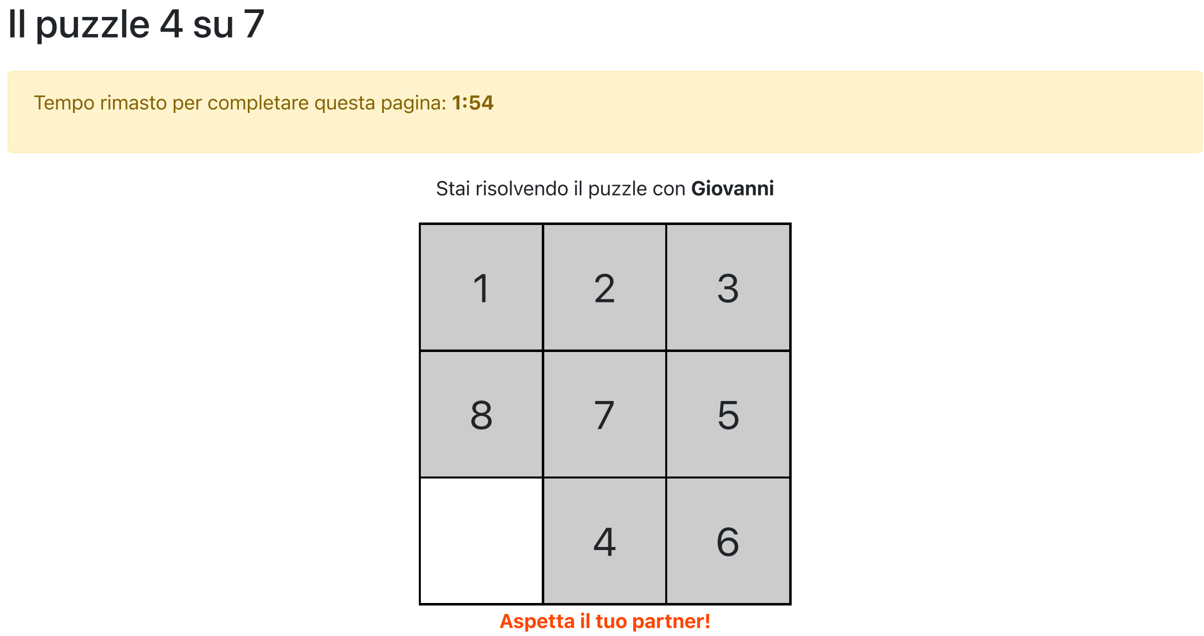
Part 2 contains seven rounds and participants learn the rules of part 3 before starting part 2. This part is based on Fisman et al. (2006, 2008)'s speed dating experiments and proceeds as follows: first, participants are allocated to a group of 8 based on their ability similarity as measured in part 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.[7] Each participant's performances in a given puzzle are measured as defined in Appendix A. 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.

6. 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.

7. Solving the puzzle itself is not incentivized, and thus participants who do not want to work 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 partner selection, participants' *intention* to correct that does not end up as an actual correction does not confound the analysis.

Figure 3: Puzzle screen

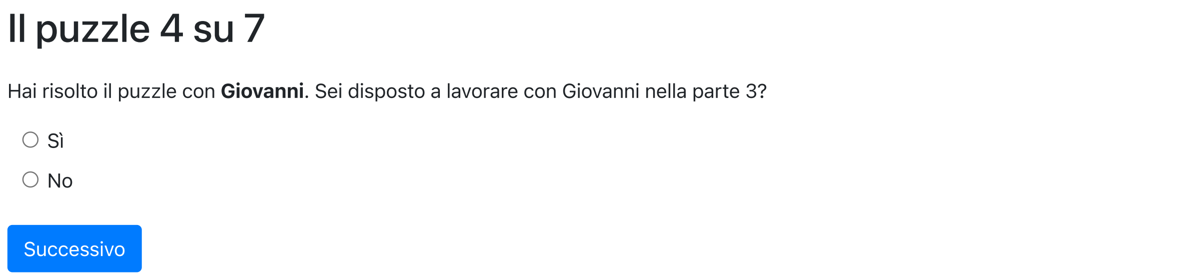


Notes: This figure shows a sample puzzle screen where a participant is matched with another participant called Giovanni at the 4th round puzzle and waiting for Giovanni to make his move.

Once they finish the puzzle, participants state whether they would like to be paired again 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 be paired again with the same participant in part 3. The sequence of puzzles is the same for all pairs in all sessions. The puzzle difficulty is kept the same across the seven rounds. The minimum number of moves to solve the puzzles is set to 8 based on the pilot.

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

Figure 4: Partner selection screen



Notes: This figure shows a sample partner 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 be paired again 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 be paired again 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 partner

The paired participants work together on the puzzles by alternating their move 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 B 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.[8]

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 on 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).[9] 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

8. None has anticipated that the puzzle is about gender.
9. The laboratory prohibits deception, so no participant has participated in an experiment with deception.

was determined by a power simulation in the pre-analysis plan to achieve 80% power.[10] The experiment and gender-related hypotheses are pre-registered with the OSF.[11]

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

3.1 Sample restrictions

I restrict the sample to puzzles where participants are paired with male participants unless otherwise indicated. This is because I am interested in men's reaction to women's correction in selecting their partner. In other words, I only use female-male and male-male pairs.

I also use part 2 data only unless otherwise indicated. This is because it is part 2 where we can observe partner selection decisions. I aggregate the move-level data at each puzzle because partner selection is observed only at each puzzle.

I use both unsolved and solved puzzles because whether a pair can solve a puzzle is an outcome of that pair. However, in the robustness check, I show that my results are robust to restricting the sample to solved puzzles only.

3.2 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.[12] 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).

3.3 Move-level summary

Figure 5 shows average move quality along with 95% confidence intervals (panel A), number of observations in each move (panel B), and number of corrections in each move (panel C), separately for female-male pairs (gray) and male-male pairs (white).

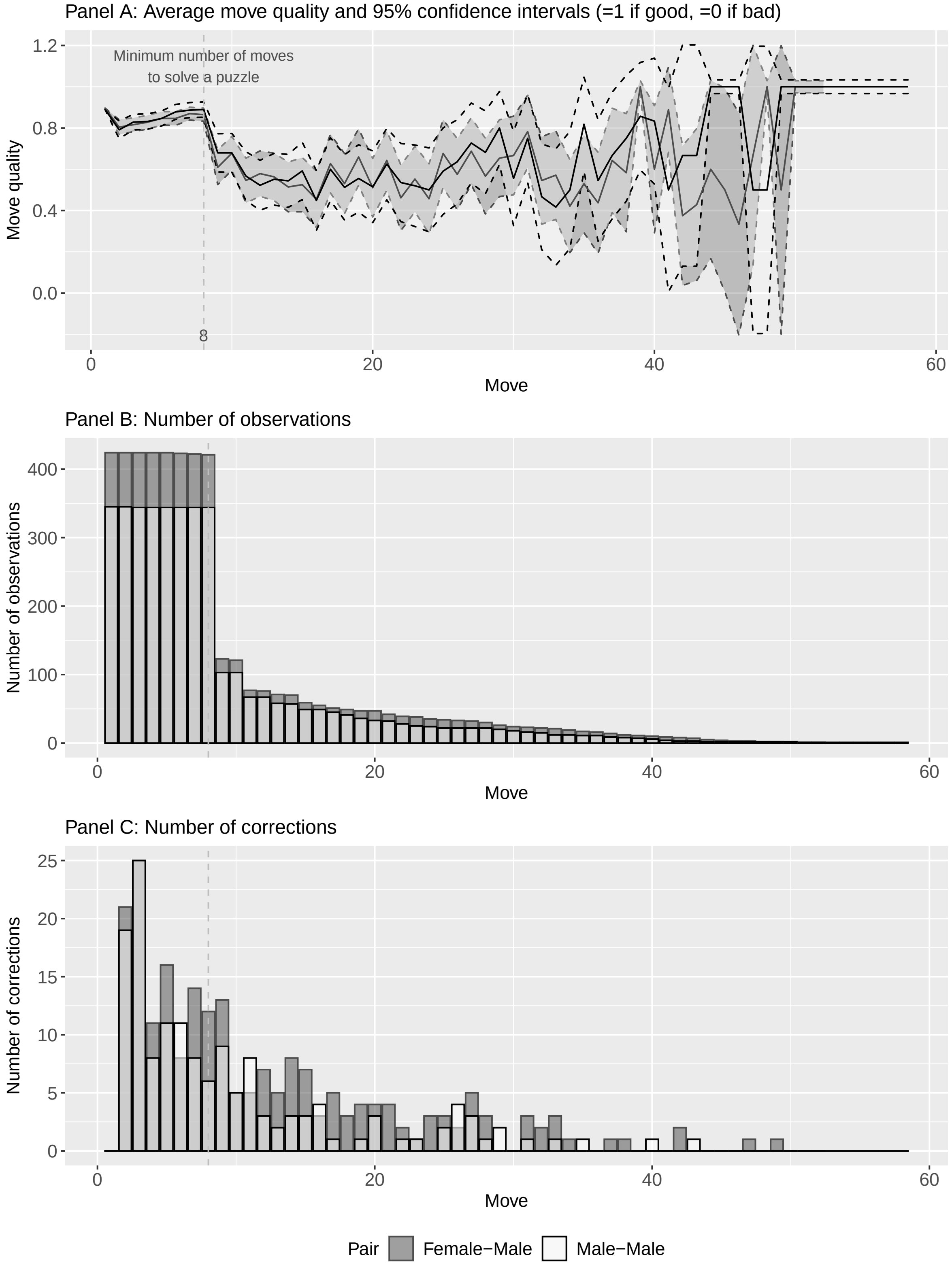
Panel A shows that the average move quality is around 0.8 (8 out of 10 are good moves) until the 8th move (the minimum number of moves to solve a puzzle). After the 8th move, move quality deteriorates and stays around 0.6 (6 out of 10 are good moves). Yet, there are no differences in the move quality between female-male pairs and male-male pairs. Panel B shows there are a little more than 400 observations up to the 8th move, then the number of observations

10. 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.

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

12. 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.

Figure 5: Move quality, number of observations, and number of corrections



Notes: The average move quality along with 95% confidence intervals (panel A), the number of observations in each move (panel B), and the number of corrections in each move (panel C), separately for female-male pairs (gray) and male-male pairs (white). 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 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 CR0 and clustered at pair level. I restrict the sample to puzzles where participants are matched with male group members.

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 B. P-values of the difference between female and male participants are calculated with HC0 heteroskedasticity-robust standard errors.

drop gradually afterward. So about 70% of the puzzles are solved with the minimum number of moves. Panel C shows that corrections happen across the moves (but not in the 1st move, by definition).

In the following, I aggregate this data at each puzzle level so that I can associate corrections, gender, and partner selection.

3.4 Puzzle-level summary

Table 2 describes own (panel A) and group members' puzzle-solving ability (panel B, the group member is always male), corrections to group member (panel C), and puzzle outcomes (panel D). Panel A shows that female participants solve 0.6 fewer puzzles in part 1. However, there are no gender differences in performance in part 2: in terms of contribution, unconstrained contribution, and net good moves. This is likely because I grouped participants with similar abilities. I elaborate on this point later in figure 7.

Panel B shows that group members' (who are all male) puzzle performance in part 2 is not different when they are paired with female or male participants. The difference in the number of puzzles solved in part 1 is statistically significant but is quantitatively negligible.

Panel C shows that participants correct group members in 15% of the total puzzles, of which 11-12% are good corrections (corrections of group members' bad moves) and 4-5% are bad corrections (corrections of group members' good moves). Also, in some puzzles, participants make corrections more than once and those are mostly good corrections. There are no gender

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

	Female (N=838)		Male (N=702)		Difference (Female – Male)		
	Mean	SD	Mean	SD	Mean	SE	P-value
Panel A: Own puzzle-solving ability							
Contribution	0.45	0.18	0.45	0.17	-0.01	0.01	0.38
# puzzles solved in pt. 1	8.30	2.42	8.88	2.40	-0.58	0.14	0.00
Contribution (unconstrained)	0.51	0.25	0.50	0.18	0.01	0.01	0.53
Net good moves	3.01	2.99	3.15	2.73	-0.14	0.16	0.36
Panel B: Group member's puzzle-solving ability (always male)							
Contribution	0.45	0.18	0.45	0.17	0.00	0.01	0.77
# puzzles solved in pt. 1	8.74	2.28	8.88	2.40	0.01	0.00	0.00
Contribution (unconstrained)	0.49	0.25	0.50	0.18	-0.01	0.01	0.53
Net good moves	3.14	2.57	3.15	2.73	-0.02	0.14	0.92
Panel C: Corrections to group member							
Correction	0.22	0.63	0.22	0.62	0.01	0.03	0.82
Good correction	0.16	0.55	0.16	0.49	0.00	0.03	0.90
Bad correction	0.06	0.28	0.06	0.32	0.00	0.02	0.80
Correction (0/1)	0.15	0.36	0.15	0.36	0.00	0.02	0.98
Good correction (0/1)	0.11	0.32	0.12	0.32	-0.01	0.02	0.74
Bad correction (0/1)	0.05	0.23	0.04	0.21	0.01	0.01	0.37
Panel D: Puzzle outcomes							
Selected as a partner (0/1)	0.71	0.45	0.70	0.46	0.01	0.03	0.69
Selected as a partner (residualized)	0.00	0.42	0.00	0.42	0.00	0.02	0.92
Time spent (sec.)	43.50	36.04	42.39	35.43	1.12	1.88	0.55
Total moves	11.28	7.88	11.12	7.49	0.16	0.43	0.70
Puzzle solved (0/1)	0.86	0.35	0.86	0.34	0.00	0.02	0.83
Consecutive correction (0/1)	0.05	0.21	0.04	0.20	0.00	0.01	0.88

Notes: This table describes own (panel A) and group member's puzzle-solving ability (panel B), corrections one made to group member (panel C), and puzzle outcomes (panel D). P-values of the difference between female and male participants are calculated with CR0 standard errors clustered at the group member level. I restrict the sample to puzzles where participants are matched with male group members. Appendix A provides definitions of each puzzle-solving ability measure.

differences in the propensity of correction.[13]

Panel D shows that participants are selected as a partner by group members 70-71% of the time. Even after netting out group member fixed effects,[14] the standard deviation of partner selection is high enough as shown in residualized partner selection, suggesting that there is enough variation in partner selection even after controlling for group member fixed effects in the analysis. Participants spend on average 42-44 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). 86% of the puzzles are solved and in 4-5% of the puzzles, participants and group members correct each other's move consecutively. There is no gender difference in any of

13. Unlike Isaksson (2018). This could be due to the presence of partner selection after the puzzle ((Isaksson 2018) does not have a partner selection stage after the puzzle).

14. Residual obtained from regressing partner selection on group member fixed effects.

these puzzle outcomes.

3.5 Gender balance and puzzle outcomes 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 6 plots average gender balance (fraction of female participants, panel A) and puzzle outcomes (panels B-J) across seven rounds along with their 95% confidence intervals. F-statistics show whether a given outcome is different across rounds.

First, panel A shows that gender is roughly balanced across rounds. Second, panels B-J show that most outcome variables are unbalanced across rounds; specifically, whether a participant is selected as a partner 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 partner 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. Later, I show that the results are robust to exclusion of rounds 6 and 7; if anything, the results get stronger (albeit with larger standard error due to reduction of the number of observations) by excluding 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 group member j who maximizes his (j is always male) expected utility by selecting their partner i from a set of j's potential partners $I \in \{1,2,3,4,5,6,7\}$. j's utility depends on his payoff and emotion. The utility is increasing in the payoff and the payoff is increasing in j's belief about i's ability. Thus, if group member j would select with whom to play in part 3, he would face the following problem:

$$\max_{i \in I} E_{\mu_i}[u_j(\underbrace{\pi(\mu_i(\tilde{a}_i, c_i, f_i))}_{\text{j's payoff}}, \underbrace{\kappa_j(c_i, f_i)}_{\text{j's emotion}})|\theta_j, \omega_j], \quad \partial u_j/\partial \pi > 0,\ \partial \pi/\partial \mu_i > 0 \tag{1}$$

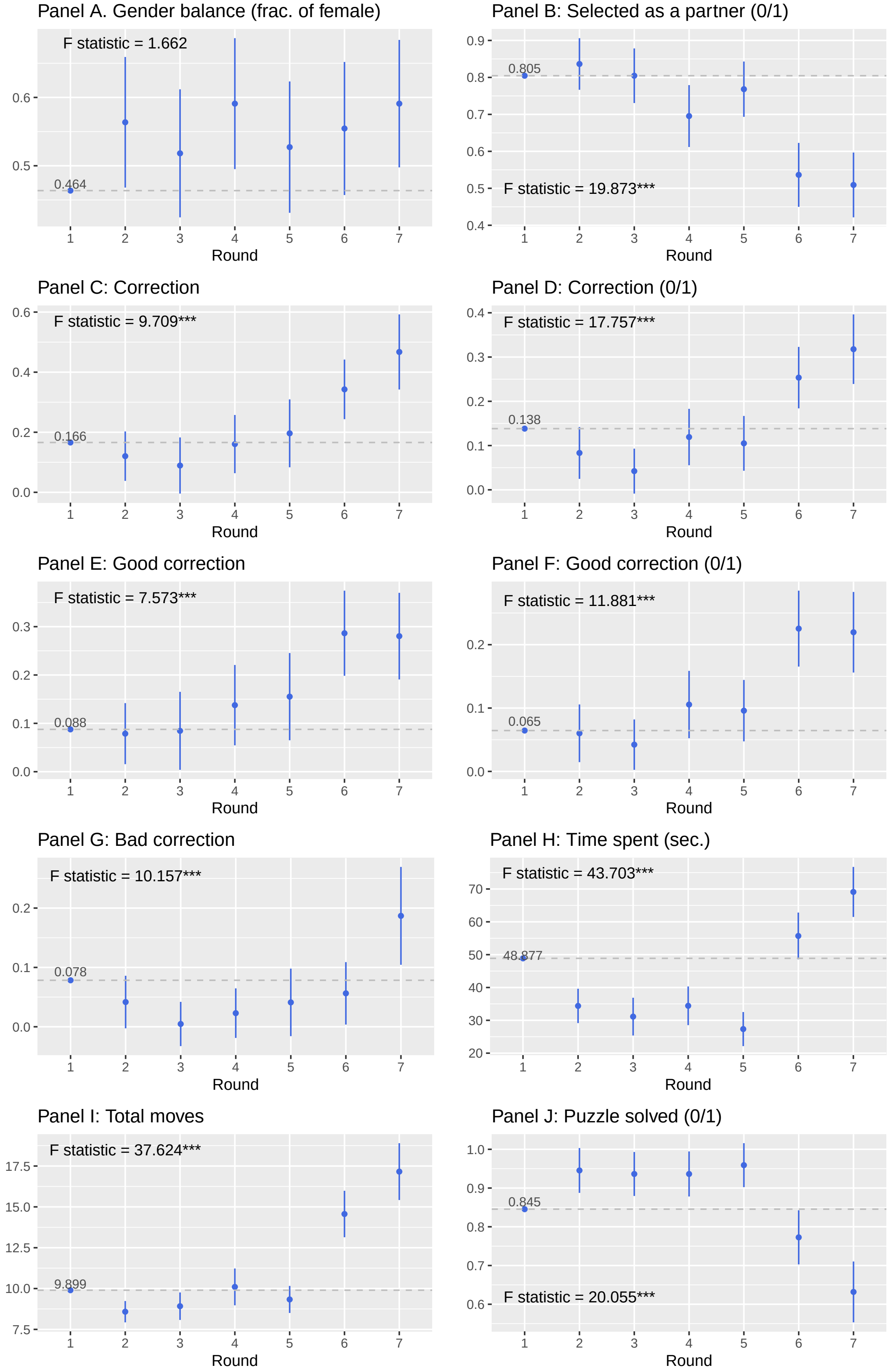
where each term is defined as follows:

- μ_i: j's belief about i's ability
- $\tilde{a}_i$: i's ability perceived by j
- c_i: i's correction (=1 if i corrected i, =0 not corrected)
- f_i: i's gender (=1 if female, =0 if male)
- θ_i: j's belief about his own ability relative to other participants (>0 if high, =0 if same, <0 if low)
- ω_i: j's belief about women's ability relative to men (>0 if high, =0 if same, <0 if low)

I make the following assumptions:

- μ_i is increasing in i's ability perceived by j: $\partial \mu_i/\partial \tilde{a}_i > 0$
- j's utility is decreasing in his emotion: $\partial u_j/\partial \kappa_j < 0$

Figure 6: Average gender balance and puzzle outcomes 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 outcomes: $y_{ij} = \beta_1 + \sum_{k=2}^{7} \beta_k \mathbb{1}[t_{ij} = k] + \epsilon_{ij}$, 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. F-statistic tests joint hypothesis under the null of $\beta_1 = \beta_k$ (k=2,...,7). The number above the 1st puzzle estimate is the 1st puzzle mean value of y_{ij}. CR0 standard errors are clustered at the group member level. I restrict the sample to puzzles where participants are matched with male group members. Significance levels: * 10%, ** 5%, and *** 1%.

- emotion is irrelevant if j is fully rational: $u_j(\pi, \kappa_j) \propto u_j(\pi)$

Because j can only partially observe i's ability, i's correction, c_i, gender, f_i, and their interaction, $c_i \times f_i$, also convey information about i's ability even if j is fully rational.

4.1 When j is fully rational

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

- $\partial\mu_i/\partial c_i < 0$ if $\theta_j > 0$,
- $\partial\mu_i/\partial c_i = 0$ if $\theta_j = 0$, and
- $\partial\mu_i/\partial c_i > 0$ if $\theta_j < 0$.

The information i's gender conveys Second, the information i's gender conveys depends on ω_j. If j believes that women's ability is high, he would consider i being a woman as a signal of high ability. On the other hand, if j believes that women's ability is low, then he would consider i being a woman as a signal of low ability. If j believes that women's ability is the same as men's, then i being a woman would not convey any information. Thus,[15]

- $\partial\mu_i/\partial f_i > 0$ if $\omega_j > 0$,
- $\partial\mu_i/\partial f_i = 0$ if $\omega_j = 0$, and
- $\partial\mu_i/\partial f_i < 0$ if $\omega_j < 0$.

The information i's correction conveys when i is a woman relative to when i is a man Last, the information i's correction when i is a woman relative to when i is a man conveys depends on ω_j; it does not depend on θ_j because it is relative to men's correction. The role of ω_j is the same as i's gender discussed above. Thus,

- $\partial^2\mu_i/\partial c_i \partial f_i > 0 \; \forall \theta_j$ if $\omega_j > 0$,
- $\partial^2\mu_i/\partial c_i \partial f_i > 0 \; \forall \theta_j$ if $\omega_j = 0$, and
- $\partial^2\mu_i/\partial c_i \partial f_i < 0 \; \forall \theta_j$ if $\omega_j < 0$.

4.2 When j is not fully rational

When j is not fully rational, j's emotion, κ_j, matters for his maximization problem. Specifically, I assume the following:

- i's correction induces j's negative feeling towards i: $\partial\kappa_j/\partial c_i < 0$
- i's correction when i is a woman induces j's stronger negative feeling towards i: $\partial^2\kappa_j/\partial c_i \partial f_i < 0$

Both assumptions are 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

15. In the analysis, I nonparametrically control for i's gender so i's gender itself does not matter in the results; what may matter is its interaction with corrections.

as biased (Kennedy and Pronin 2008). The second assumption is based on the finding that men are motivated stereotyper: men evaluate women in a stereotypical way when those women criticize them (Sinclair and Kunda 2000). While they are both beliefs, I assume they affect j's actions. We can also base the second assumption on in-group/out-group bias (Tajfel and Turner 1979): when a person who belongs to an out-group (a group of people who do not share the same identity as oneself, in this case, women) does something bad to us, we react to it more negatively. This is similar to Chen and Li (2009)'s finding that people punish out-group members' misbehavior more.

4.3 Questions to test and rational agent benchmark

I provide the following questions that navigate interpretations of the results and predictions for a rational agent derived from the theoretical framework; in all questions and predictions, I fix i's ability observed by j:

Question 1. *Do men believe women and men are equally good at the puzzle; that is, is $\omega_j = 0$? — **Rational prediction**: Unclear, but likely to be yes since we already see that there is no gender difference in the puzzle ability and that Isaksson (2018) finds the same.*

Question 2. *Are men less likely to select as a partner a person who corrected their move; that is, is $\partial\mu_i/\partial c_i < 0$? — **Rational prediction**: Unclear; it depends on j's belief about his ability relative to i.*

Question 3. *Are men less likely to select as a partner a woman than a man who corrected their move; that is, is $\partial^2\mu_i/\partial c_i \partial f_i < 0$? — **Rational prediction**: No, as long as the answer to question 1 is yes.*

Remember that there are good and bad corrections and that group members can partially observe the quality of each move. Thus, although not modeled explicitly, good corrections must convey more positive information about i than bad corrections. Thus, I also test the following question.

Question 4. *For both questions 2 and 3, does a correction that corrected men's wrong move receive less negative/more positive reaction? — **Rational prediction**: Yes, regardless of self-confidence and gender bias, participants should consider less negatively/more positively those who corrected their bad move as a partner.*

5 Empirical strategy

Keeping participant's perceived ability fixed, I need to observe group member's partner selection in the four conditions shown in table 3 to answer questions in section 4.3 where participant's gender and correction are exogenously varied and the group member is always male.

By random pairing of participants, the participant's gender is exogenous to group member's unobservables. However, correction is not exogenous for two reasons:

1. Correction can be correlated with the participant's ability and participant's ability can affect group member's partner selection.

Table 3: Conditions under which I need to observe group member's partner selection to answer questions in section 4.3

		Participant's gender	
		Female	Male
Corrected group member (always male)	Yes	A	B
	No	C	D

Notes: This table shows conditions under which I need to observe group member's partner selection to answer questions in section 4.3, keeping participant's perceived ability fixed. Participant's gender and correction must be exogenously varied. The group member must always be male.

2. There is an effect similar to the reflection effect: group member's puzzle behavior affects the participant's behavior and vice versa; for example, a group member's meanness can increase the participant's correction and can also affect his partner selection.

To address the first point, I control for participant's ability both by design and econometrically.[16] To address the second point, I include group member fixed effects. Under the assumption that the participant's ability *perceived by the group member* is fully controlled for, the sources of variation in correction are participant's meanness, random move, perceived puzzle difficulty, etc., which are orthogonal to the group member's unobservables.

5.1 Ability measure selection

Figure 7 shows the empirical distributions of four ability measures (see Appendix A for the definition of each measure): panels A-D include both solved and unsolved puzzles and panels E-H include solved puzzles only. While there are four ability measures, the figure shows that the number of puzzles solved in part 1 (panels B and F) is not an appropriate measure because it does not seem to represent participant's ability in part 2 well – all the ability measures from part 2 do not resemble the number of puzzles solved in part 1. While unconstrained contribution and net good moves both capture participant's ability in part 2, they have a long left tail. In addition, group members will likely consider participants whose ability is really bad as equally unsuitable as a partner. These points are elaborated on in Appendix C.

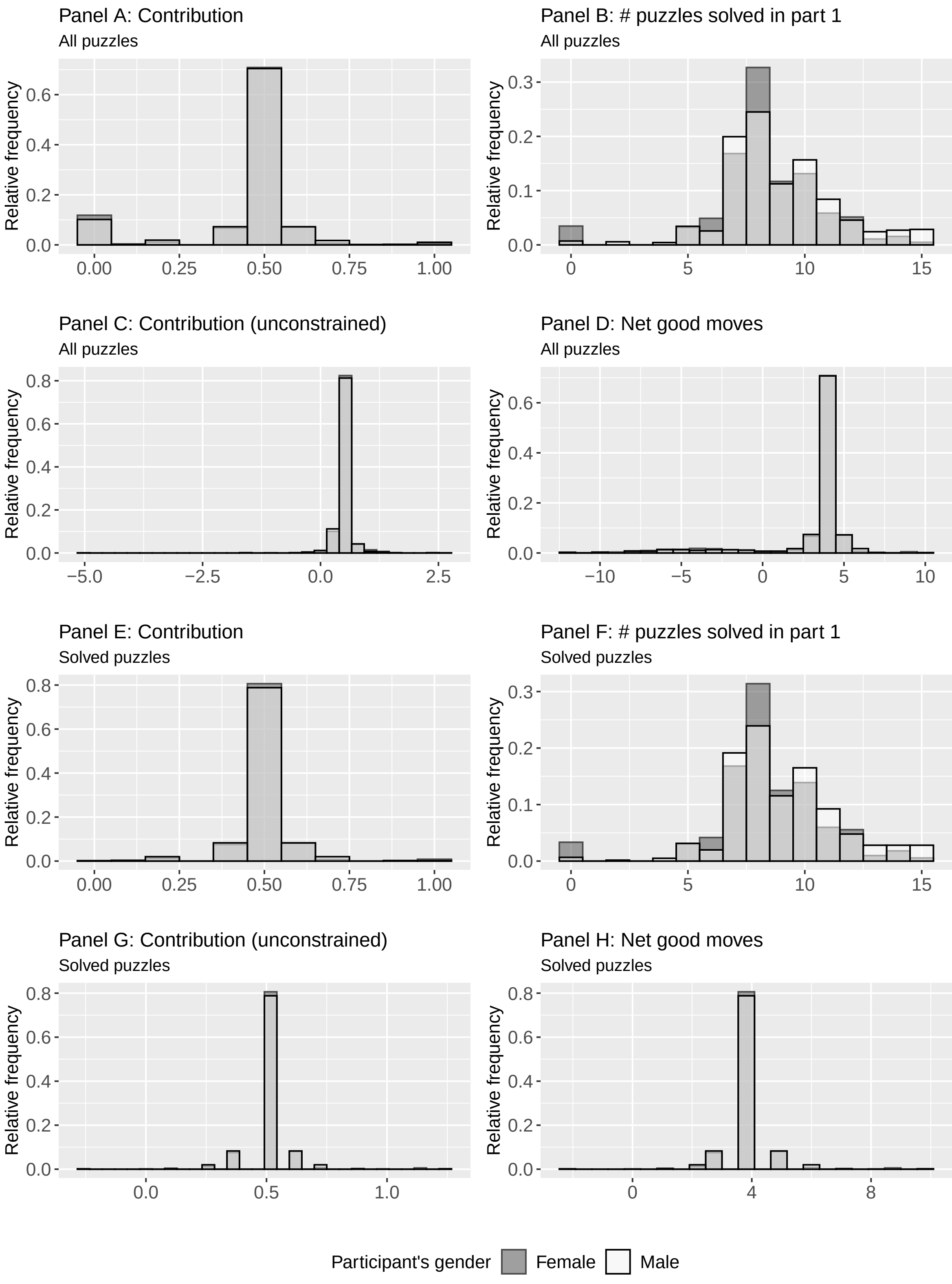
Thus, I use contribution as a measure of ability. Because I grouped participants with similar abilities, both participants contributed equally to more than 70% of puzzles, which makes it more convincing that controlling for contribution econometrically would capture almost all the *perceived* (not true) ability of the participant.

5.2 Estimating equation

I run the following OLS regression. As discussed earlier, I restrict the sample to puzzles where participants are paired with male participants: I only use female-male or male-male pairs (so i

16. I could vary whether participants can correct group member, but it alters the rule of the puzzle, which alters participants' ability.

Figure 7: Gender differences in puzzle-solving ability



Notes: This figure shows the distribution of ability measures separately for female (gray) and male (white) participants. Panels A-D shows the distributions for both solved and unsolved puzzles and panels E-H for solved puzzles only. Appendix A provides definitions of each ability measure.

can be either female or male but j is always male).

$$Selected_{ij} = \beta_0 + \beta_1 Correct_{ij} \times Female_j + \beta_2 Correct_{ij} + \beta_3 Female_j + \delta Contribute_{ij} + \mu_j + \epsilon_{ij}$$
$$i \in \{Female, Male\}, j \in \{Male\} \tag{2}$$

where each variable is defined as follows:

- $Selected_{ij} \in \{0, 1\}$: an indicator variable equals 1 if i is selected by j as their partner, 0 otherwise.
- $Correct_{ij} \in \{0, 1, ...\}$: the number of times i corrects j's move.
- $Female_i \in \{0, 1\}$: an indicator variable equals 1 if i is female, 0 otherwise.
- $Contribute_{ij} \in [0, 1]$: i's contribution to a puzzle played with j.
- ϵ_{ij}: omitted factors that affect i's likelihood to be selected by j as their partner.

and $\mu_j \equiv \sum_{k=1}^{N} \mu^k \mathbb{1}[j = k]$ is group member fixed effects, where N is the total number of group members in the sample and $\mathbb{1}$ is the indicator variable. Standard errors are clustered at the group member level.[17]

Under the assumption that contribution almost fully controls for i's ability observed by j, coefficients correspond to table 3 as follows:

Table 4: Conditions equation 2's coefficients identify

		Participant's gender	
		Female	Male
Corrected group member (always male)	Yes	$\beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_0 + \beta_2$
	No	$\beta_0 + \beta_3$	β_0

Notes: This table shows conditions equation 2's coefficients identify, keeping i's perceived ability fixed. Group members are always male.

Although I cannot estimate the intercept term, β_0, because of group member fixed effects, I can still test the questions set up in section 4.3:

- $\beta_3 = (\beta_0 + \beta_3) - \beta_0 = \partial u_j / \partial f_i |_{c_i=0}$ (some evidence for question 1)
- $\beta_2 = (\beta_0 + \beta_2) - \beta_0 = \partial u_j / \partial c_i$ ($= \partial \kappa_j / \partial c_i$ if $\theta_j = 0$) (test of question 2)
- $\beta_1 = [(\beta_0 + \beta_1 + \beta_2 + \beta_3) - (\beta_0 + \beta_3)] - [(\beta_0 + \beta_2) - \beta_0] = \partial^2 u_j / \partial c_i \partial f_i$ ($= \partial^2 \kappa_j / \partial c_j \partial f_i$ if $\omega_j = 0$) (test of question 3)

6 Results

Column 3 of Table 5 presents regression results of equation 2. Columns 1 and 2 omit some controls to show direction and magnitude of omitted variable bias, column 4 shows whether the number of puzzles solved in part 1 matters in group members' partner selection and column 5 shows whether a group member with higher gender bias responds more negatively/less positively to women's

17. This is because the treatment unit is j. Although the same participant appears twice (once as i and once as j), i is passive in preference elicitation.

Table 5: Cost of correcting male group member's move

Outcome:	Selected as a partner by the group member (0/1)				
Sample:	Puzzles solved with male group member				
	(1)	(2)	(3)	(4)	(5)
Correct×Female	-0.017	-0.027	-0.003	-0.003	-0.005
	(0.044)	(0.045)	(0.044)	(0.044)	(0.063)
Correct	-0.149***	-0.152***	-0.122***	-0.123***	-0.139**
	(0.025)	(0.028)	(0.040)	(0.040)	(0.055)
Female	0.017	0.013	0.016	0.020	0.019
	(0.024)	(0.026)	(0.022)	(0.023)	(0.030)
Contribution			1.222***	1.218***	1.224***
			(0.068)	(0.068)	(0.069)
# puzzles solved alone				0.008	
				(0.006)	
Group member high bias					0.012
×Correct×Female					(0.086)
Group member high bias					0.028
×Correct					(0.079)
Group member high bias					-0.012
×Female					(0.045)
Group member FE	-	✓	✓	✓	✓
Correct×Female	-0.166***	-0.179***	-0.125***	-0.126***	-0.143***
+Correct	(0.035)	(0.036)	(0.025)	(0.025)	(0.042)
Group member high bias					0.007
×Correct×Female					(0.059)
+Correct×Female					
Baseline mean	0.747	0.747	0.747	0.747	0.749
Baseline SD	0.435	0.435	0.435	0.435	0.434
Adj. R-squared	0.047	0.061	0.299	0.300	0.298
Observations	1510	1510	1510	1510	1503
Clusters	220	220	220	220	219

Notes: This table presents regression results of equation 2. Column 1 does not control for contribution to the puzzle and does not have group member fixed effects, column 2 controls for contribution to the puzzle but does not have group member fixed effects, and column 3 controls for contribution to the puzzle and has group member fixed effects. Column 4 additionally controls for the number of puzzles solved in part 1 to show it does not alter the results in column 3. Column 5 separates the coefficient estimates for male group members with higher and lower gender bias to see whether male group members with higher bias respond more negatively/less positively to women's corrections. Baseline mean and standard deviation are that of men who do not correct group members. CR0 standard errors in parentheses are clustered at the group member level. Significance levels: * 10%, ** 5%, and *** 1%.

corrections. The most important thing to note is that a matched participant's contribution is the main determinant of a male group member's partner selection, as the coefficient estimate on contribution is statistically and quantitatively highly significant. This suggests that the partner selection is well-incentivized and that group members can partially observe matched participants' abilities.

Second, looking at columns 1 and 2, coefficient estimates on correction (β_2 in equation 2), its interaction with female dummy (β_1 in equation 2), and their sum ($\beta_1 + \beta_2$ in equation 2) are more negative than those in column 3. Other coefficient estimates are similar to column 3. These suggest that not controlling for contribution makes the correction effect stronger. Comparing columns 1 and 2, not adding group member fixed effects seem to make the correction effect slightly weaker, especially for women's correction.

Third, column 4 adds the number of puzzles solved in part 1 by the participant as an additional puzzle ability control to see whether part 1 behavior affects the group member's partner selection. However, compared to column 3, adding the number of puzzles solved in part 1 changes almost nothing: all coefficient estimates stay the same, and R-squared only increases by 0.001. This makes us more confident that while there are some gender differences in part 1, it is irrelevant for group member's partner selection.

Below, I discuss how column 3 answers each of the questions I posed in section 4 below. I also discuss what column 5 tells us about whether a group member with higher gender bias responds more negatively/less positively to women's corrections.

6.1 Do men believe women and men are equally good at the puzzle? – Some evidence for question 1

The coefficient estimate on the female dummy of column 3 of table 5 (β_3 in equation 2) is the probability that male group members select women over men as their partners in absence of corrections. As we see in column 3, it is statistically and quantitatively insignificant; in fact, it is slightly positive. Thus, absence of correction, men are not less likely to select women over men as their partners.

Although this result does *not* evidence that men believe women and men are equally good at the puzzle, there is no reason for a group member to select a participant whom they believe has a lower ability. Also, as we saw in table 2 and figure 7, there is no gender difference in puzzle ability in part 2 both in mean and distribution: although men solved slightly more puzzles in part 1, participants do not observe other participants' part 1 behavior. Further, Isaksson (2018) who uses the same puzzle also finds no gender difference in puzzle ability. Thus, taken together, it is likely that men believe women and men are equally good at the puzzle – at least statistical discrimination story is unlikely as an explanation of any gender differences I would find in the analysis.

6.2 Are men less likely to select as a partner a person who corrected their move? – Formal test of question 2

The coefficient estimate on correction of column 3 of Table 5 (β_2 in equation 2) is the effect of a male participant's correction on a male group member's probability to select that participant as a partner. It is negative and statistically and quantitatively highly significant. This suggests that correcting a male group member reduces the probability of being selected into teamwork by 12.2 percentage points. Relative to the baseline mean (the male participant's probability to be selected by a male group member as a partner in absence of correction), the effect is 16.3%.

Remember that this effect is causal, that it is comparing the two participants who contributed equally to the puzzle but one corrected the group member and the other did not. So how much does one have to increase their contribution in order to offset the negative effect of correction? Using the standard deviation of men's contribution in table 2 and coefficient estimate of contribution in column 3 of table 5, back of the envelope calculation shows: $|\hat{\beta}_2|/(\hat{\delta} \times SD_{contribution,male}) = |-0.122|/(1.222 \times 0.17) \approx 0.59$. So one has to increase their contribution by 0.59 standard deviation to be equally preferred by the group member as a partner, which is pretty large.

6.3 Are men less likely to select as a partner a woman than a man who both corrected their move? – Formal test of question 3

The coefficient estimate on correction times female dummy of column 3 of table 5 (β_1 in equation 2) is the effect of female participant's correction relative to male participant's correction on male group member's probability of select her as a partner. Although it is negative, it is neither statistically nor quantitatively significant. This suggests that women's correction does not receive a stronger negative reaction relative to men's.

Nevertheless, the total effect of women's correction on the probability to be selected as a partner by male group members – the sum of coefficient estimates on correction times female dummy and on correction ($\beta_1 + \beta_2$ in equation 2) – is negative and statistically and quantitatively highly significant. Thus, even though women's correction does not receive a stronger reaction, this result suggests that correcting male group members is costly also for women.

Heterogeneity by men's gender bias Male group members with high gender bias may consider women as lower ability than men or react more negatively to women's correction. Thus, I test this possibility by separating the effect of correction for male group members with higher and weaker gender bias, which I do by augmenting equation 2 and re-estimating it via OLS:

$$\begin{aligned} Selected_{ij} = &\beta_0 + \beta_1 Correct_{ij} \times Female_i + \beta_2 Correct_{ij} + \beta_3 Female_i \\ &+ \beta_4 Correct_{ij} \times Female_i \times HighBias_j + \beta_5 Correct_{ij} \times HighBias_j \\ &+ \beta_6 Female_i \times HighBias_j + \delta Contribute_{ij} + \mu_j + \epsilon_{ij} \\ &\qquad i \in \{Female, Male\}, j \in \{Male\} \end{aligned} \tag{3}$$

where each variable is defined as follows:

- $HighBias_j \in \{0,1\}$: an indicator variable equals 1 if j's gender bias is above the median of all male participants, 0 otherwise.

other variables are defined as in equation 2.

The results are presented in column 5 of Table 5 and the terms interacted with "Group member high bias" show the difference of the effect of those terms for male group members with higher and lower gender bias. In short, I do not find any statistically or quantitatively significant difference for any effect. Although, this could be due to that participants did not answer the gender bias questions honestly because it is a socially sensitive issue.

Table 6: Cost of correcting male group member's bad move

Outcome:	Selected as a partner by the group member (0/1)			
Sample:	Puzzles solved with male group member			
	(1)	(2)	(3)	(4)
Correct×Female	-0.115	-0.115	0.035	0.037
	(0.076)	(0.082)	(0.069)	(0.069)
Correct	-0.276***	-0.267***	-0.031	-0.033
	(0.051)	(0.052)	(0.046)	(0.047)
Female	0.023	0.018	0.013	0.016
	(0.024)	(0.026)	(0.022)	(0.023)
CorrectGood×Female	0.103	0.091	-0.029	-0.033
	(0.098)	(0.104)	(0.079)	(0.079)
CorrectGood	0.189***	0.182***	-0.140**	-0.137**
	(0.064)	(0.064)	(0.061)	(0.062)
Contribution			1.304***	1.300***
			(0.077)	(0.077)
# puzzles solved alone				0.007
				(0.006)
Group member FE	-	✓	✓	✓
Correct×Female	-0.392***	-0.382***	0.004	0.004
+Correct	(0.055)	(0.061)	(0.055)	(0.055)
CorrectGood×Female	0.291***	0.273***	-0.169***	-0.170***
+CorrectGood	(0.073)	(0.078)	(0.066)	(0.065)
Baseline mean	0.747	0.747	0.747	0.747
Baseline SD	0.435	0.435	0.435	0.435
Adj. R-squared	0.064	0.075	0.304	0.304
Observations	1510	1510	1510	1510
Clusters	220	220	220	220

Notes: This table presents regression results of equation 4. Column 1 does not control for contribution to the puzzle and does not have group member fixed effects, column 2 controls for contribution to the puzzle but does not have group member fixed effects, and column 3 controls for contribution to the puzzle and has group member fixed effects. Column 4 additionally controls for the number of puzzles solved in part 1 to show it does not alter the results in column 3. Baseline mean and standard deviation are that of men who do not correct group members. CR0 standard errors in parentheses are clustered at the group member level. Significance levels: * 10%, ** 5%, and *** 1%.

6.4 Does a correction that corrected a men's wrong move receive less negative/more positive reaction? – Formal test of question 4

So far I find men are reluctant to select as a partner those who corrected their move, but it does not reduce group efficiency if their reluctance only concerns corrections that correct their good move. To test whether this is the case, I separate the effect of good and bad correction, which I do by augmenting equation 2 and re-estimating it via OLS:

$$
\begin{aligned}
Selected_{ij} = \beta_0 + \beta_1 Correct_{ij} \times Female_i + \beta_2 Correct_{ij} + \beta_3 Female_i + \beta_4 CorrectGood_{ij} \times Female_i \\
+ \beta_5 CorrectGood_{ij} + \delta Contribute_{ij} + \mu_j + \epsilon_{ij} \\
i \in \{Female, Male\}, j \in \{Male\}
\end{aligned}
\tag{4}
$$

where each variable is defined as follows:

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

other variables are as defined in equation 2.

The results are presented in column 3 of Table 6. As table 5, columns 1 and 2 show bias arising from omitting contribution and group member fixed effects, and column 4 adds the number of puzzles solved in part 1 as an additional puzzle ability control to show part 1 behavior does not affect group members' partner selection.

The coefficient estimate on correction of column 3 of table 6 (β_2 in equation 4) is the effect of a male participant's good correction on the male group member's probability to select that participant as a partner relative to a bad correction. It is negative, quantitatively significant, and statistically significant at 5%. This suggests that correcting a male group member's bad move reduces the probability of being selected into teamwork more than correcting a male group member's good move.

It is important to note that I am comparing the two participants who contributed equally to the puzzle but one corrected the group member's bad move and the other corrected a good move, and the results do not mean good correction receives more negative reaction. In fact, when I do not control for puzzle ability, reported in columns 1 and 2 of table 6, good correction increases the probability of being selected as a partner and bad correction decreases the probability.

Thus, what these results suggest is that while male group members appreciate the contribution part of the correction, they dislike the part that points out their mistakes, hence missing an opportunity to select a good partner and missing a successful teamwork opportunity.

Although there is no gender effect – good corrections by both women and men equally reduce the probability of being selected into teamwork – the sum of coefficient estimates on good correction times female dummy and on correction ($\beta_4 + \beta_5$ in equation 4) is negative and statistically and economically significant.

6.5 Why do men react negatively to corrections?

So why do male group members react negatively to corrections? One possible reason is their overconfidence as men are overconfident (Croson and Gneezy 2009). Remember that the correctness of the correction is not fully observable and corrections convey information about the matched participant's ability as we saw in section 4. Also, remember that group members are told how many puzzles they have solved in part 1 at the end of that part. Thus, those who solved fewer puzzles in part 1 should be less confident about their ability than those who solved more puzzles. This does not affect the matched participants' ability relative to them because group members are grouped with participants of similar abilities. Thus, if their negative reaction

Table 7: Mechanism of male group member's negative reaction to corrections

Outcome:	Selected as a partner by the group member (0/1)			
Sample: Puzzles solved with	male group member		female & male group member	
	(1)	(2)	(3)	(4)
Correct	-0.179***	-0.054	-0.125***	-0.028
	(0.029)	(0.034)	(0.024)	(0.034)
Female	0.014	0.014	0.009	0.009
	(0.021)	(0.021)	(0.014)	(0.014)
CorrectGood		-0.182***		-0.136***
		(0.054)		(0.044)
Contribution	1.230***	1.312***	1.199***	1.246***
	(0.068)	(0.076)	(0.050)	(0.054)
Group member low ability×Correct	0.083*	0.062		
	(0.042)	(0.057)		
Group member low ability×CorrectGood		0.041		
		(0.077)		
Group member female×Correct			-0.067**	-0.106**
			(0.030)	(0.046)
Group member female×CorrectGood				0.060
				(0.059)
Group member FE	✓	✓	✓	✓
Group member low ability×Correct +Correct	-0.096***	0.008		
	(0.031)	(0.052)		
Group member low ability×CorrectGood +CorrectGood		-0.141**		
		(0.064)		
Group member female×Correct +Correct			-0.191***	-0.134***
			(0.019)	(0.036)
Group member female×CorrectGood +CorrectGood				-0.076*
				(0.044)
Baseline mean	0.747	0.747	0.747	0.747
Baseline SD	0.435	0.435	0.435	0.435
Adj. R-squared	0.302	0.307	0.318	0.321
Observations	1510	1510	3180	3180
Clusters	220	220	464	464

Notes: This table presents regression results of equations 5 (column 1), 6 (column 2), 7 (column 3), and 8 (column 4). In all results, the interaction between correction and female dummy is dropped to increase efficiency and not to pick up an imbalance in the data. Columns 1 and 2 separate the coefficient estimates for male group members with higher and lower ability to see whether the results in column 3 of Table 5 and of table 6 are driven by their overconfidence. Columns 3 and 4 include both female and male group members (that is, $j \in \{Female, Male\}$) and examines gender differences in response to correction. Baseline mean and standard deviation are that of men who do not correct group members. CR0 standard errors in parentheses are clustered at the group member level. Significance levels: * 10%, ** 5%, and *** 1%.

is driven by their overconfidence, those who solved fewer puzzles should respond less negatively to corrections than those who solved more puzzles because they are less likely to believe their moves to be correct.

I test this conjecture by separating the effect of correction for male group members who solved fewer and more puzzles by augmenting equation 2 and re-estimating it via OLS. Because I find women's corrections do not receive a stronger negative reaction than men's, I drop interaction between correction and female dummy to increase efficiency. It is also because OLS may pick up an imbalance in the data with too many interaction terms when the data only has a moderate number of clusters such as mine.

$$Selected_{ij} = \beta_0 + \beta_1 Correct_{ij} + \beta_2 Female_i + \beta_3 Correct_{ij} \times LowAbility_j + \delta Contribute_{ij} + \mu_j + \epsilon_{ij}$$
$$i \in \{Female, Male\}, j \in \{Male\} \tag{5}$$

where each variable is defined as follows:

- $LowAbility_j \in \{0, 1\}$: an indicator variable equals 1 if j solved a below-median number of puzzles in part 1 in a given session, 0 otherwise.

other variables are defined as in equation 2.

The results are presented in column 1 of Table 7. First, the coefficient estimate on correction, which is the effect of correction on high-ability group members, is more negative than the estimate in column 3 of Table 5. Second, the coefficient estimate on the interaction between correction and low-ability group members is positive and statistically significant at 10%. These results seem to suggest that it may be male group members' overconfidence that is driving their negative reaction to corrections.

However, because low-ability group members cannot observe move quality as well as high-ability group members, the results above could simply be due to a difference in their ability rather than a difference in their overconfidence. In other words, the results above are coming from high-ability group members' negative reaction to bad correction.

To test this possibility, I separate the effect of good and bad correction for high- and low-ability group members. If the above finding comes from ability difference, then high-ability group members should respond less negatively to good correction and more negatively to bad correction. I do this by augmenting equation 4 and re-estimating it via OLS. As in equation 5, I drop interactions between good and bad correction and the female dummy for the same reason I drop them in equation 5.

$$\begin{aligned} Selected_{ij} = & \beta_0 + \beta_1 Correct_{ij} + \beta_2 Female_i + \beta_3 CorrectGood_{ij} + \beta_4 Correct_{ij} \times LowAbility_j \\ & + \beta_5 CorrectGood_{ij} \times LowAbility_j + \delta Contribute_{ij} + \mu_j + \epsilon_{ij} \\ & i \in \{Female, Male\}, j \in \{Male\} \end{aligned} \tag{6}$$

where each variable is defined as in equations 4 and 5.

The results are presented in column 2 of Table 7. First, the coefficient estimate on correction, which is the effect of bad correction on high-ability group members, is slightly more negative than the estimate in column 3 of Table 6. Second, the coefficient estimate on good correction, which is the difference between the effect of good and bad correction on high-ability group members, is more negative than the estimate in column 3. Third, the coefficient estimate on the interaction between correction and low-ability group members, which is the difference between the effect of bad correction on high and low-ability group members, is positive although statistically insignificant. Fourth, the coefficient estimate on the interaction between good correction and low-ability group members, which is a double difference between the effect of good and bad correction on high and low-ability group members, is positive although statistically insignificant. These are *inconsistent* with the story that the results we saw in column 1 of Table 6 are simply due to that high-ability group members can observe move quality better than low-ability group members.

6.6 Gender differences in responses to corrections (exploratory)

So far I focus on men's response to corrections. However, since I have data on women's responses to correction, I examine its gender differences as an exploratory analysis.

Response to correction To test gender differences in response to corrections, I pool both female and male group members and separate the effect of correction for female and male group members. Because I find women's corrections do not receive a stronger negative reaction than men's, I drop interaction between correction and female participant dummy to make interpretation easier.

$$Selected_{ij} = \beta_0 + \beta_1 Correct_{ij} + \beta_2 Female_i + \beta_3 Correct_{ij} \times Female_j + \delta Contribute_{ij} + \mu_j + \epsilon_{ij}$$
$$i \in \{Female, Male\}, j \in \{Female, Male\} \tag{7}$$

where each variable is defined as follows:

- $Female_j \in \{0, 1\}$: an indicator variable equals 1 if j is female, 0 otherwise.

other variables are defined as in equation 2.

Column 3 of Table 7 presents the results of equation 7. First, coefficient estimates on correction, female dummy, and contribution are almost the same as the estimates in column 3 of Table 5. Second, however, the coefficient estimate on the interaction between correction and female group member dummy is negative and statistically significant at 5%, suggesting that women react more negatively to correction than men. It is quantitatively significant too: correcting a female group member reduces the probability of being selected as a partner by 19.1 percentage points. Relative to the baseline mean (the male participant's probability to be selected as a partner by group members in absence of correction), the effect is 25.6%.

Response to good vs. bad correction I next test gender differences in response to good and bad corrections. As in equation 7, I pool both female and male group members and separate the effect of good and bad corrections for female and male group members. As in equation 7, I drop the interaction between correction and the female participant dummy to make the interpretation easier.

$$\begin{aligned} Selected_{ij} = & \beta_0 + \beta_1 Correct_{ij} + \beta_2 Female_i + \beta_3 CorrectGood_{ij} + \beta_4 Correct_{ij} \times Female_j \\ & + \beta_5 CorrectGood_{ij} \times Female_j + \delta Contribute_{ij} + \mu_j + \epsilon_{ij} \\ & i \in \{Female, Male\}, j \in \{Female, Male\} \end{aligned} \tag{8}$$

where each variable is defined as in equations 4 and 5.

Column 4 of Table 7 presents the results of equation 8. First, the coefficient estimate on the interaction between correction and female group member dummy is negative and statistically significant at 5%, suggesting that women respond more negatively to bad correction than men. Second, however, the coefficient estimate on the interaction between good correction and female group member dummy is positive and quantitatively significant, although statistically insignificant. Third, looking at the sum of coefficient estimates on the interaction between good correction and female group member dummy and on good correction, it is negative but less so than male group members, and statistically significant only at 10% despite that the sample size has doubled.

These results suggest that while women react more negatively to corrections, the effect mainly comes from bad correction and they respond less negatively to good corrections than men.

6.7 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.

7 Robustness checks

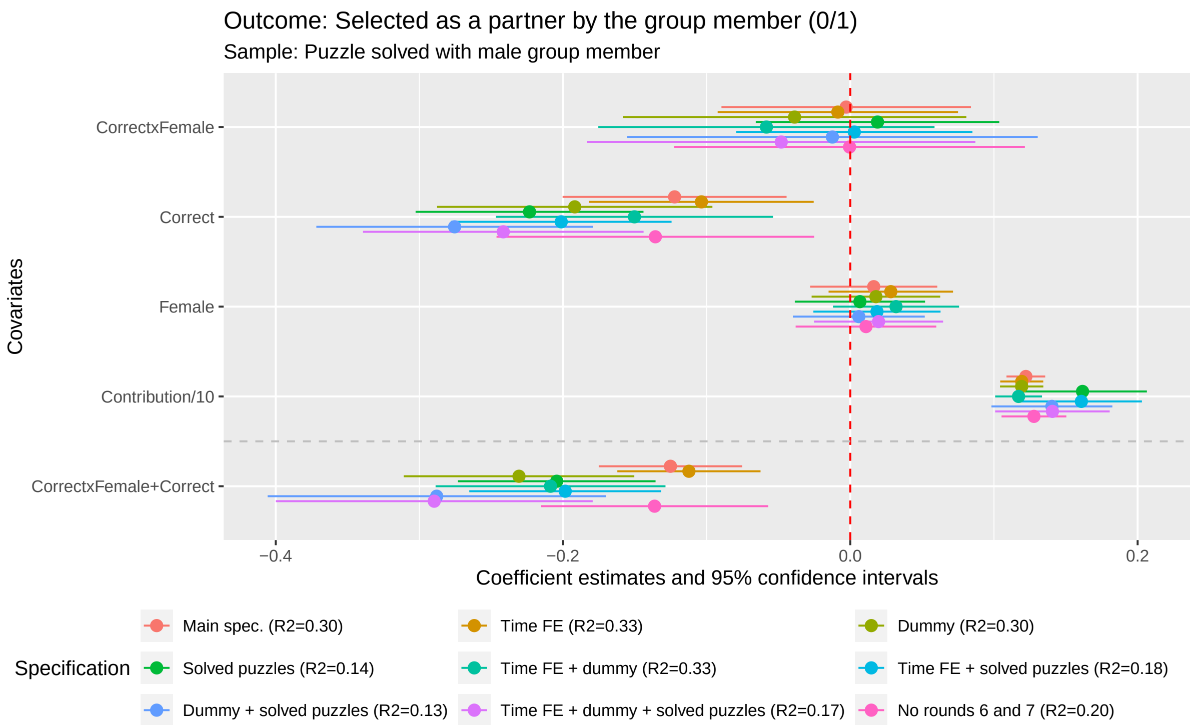
7.1 Robustness of the results for questions 1-3

In figure 8, I present several alternative specifications of equation 2 to make sure that the results for questions 1-3 I presented in column 3, Table 5, are robust to alternative explanations. The y-axis presents covariates and the x-axis presents their coefficient estimates (points) and 95% confidence intervals (lines). To facilitate comparison, I present results of column 3 of table 5 on the top in red labeled as "Main spec." Below, I raise alternative explanations one by one and explain how the figure rules out them.

Is the correlation in rounds 6 and 7 causation? We saw in figure 6 that there is a negative correlation between partner selection and correction in rounds 6 and 7, and although we control for participant's contribution to the puzzle, one may wonder if the contribution is not appropriately controlling for the observed ability and the results are capturing this correlation.

To address this concern, I re-estimate equation 2 excluding rounds 6 and 7 observations, and the results are presented in figure 8 in pink, labeled as "No rounds 6 and 7." The figure shows that this concern is invalid: all the coefficient estimates are about the same as the estimates of "Main spec." The confidence intervals are wider due to a drop in the number of observations by 2/7.

Figure 8: Cost of correcting male group member's move (robustness)



Notes: This figure shows several alternative specifications of equation 2 to show the robustness of the results of column 3, table 5 along with their adjusted R-squared. The specification "Main spec." is the column 3 results to make comparison easier. The y-axis shows covariates and the x-axis shows coefficient estimates of those covariates along with their 95% confidence intervals. The specification "Time FE" adds time fixed effects to equation 2 to show robustness to ex-post imbalance across rounds. The specification "Dummy" replaces correction in equation 2 with its dummy to show robustness to significant non-linearity. The specification "Solved puzzles" restricts the sample to solved puzzles only to show that the results are not driven by unsolved puzzles. The specification "No rounds 6 and 7" excludes observations in rounds 6 and 7 to show that the negative correlation between partner selection and correction in rounds 6 and 7 is not causation. Other specifications add combinations of these specifications. CR0 standard errors are clustered at the group member level.

Ex-post imbalance across rounds? Although ex-ante all rounds must be balanced by random pairing, one may concern that there is an ex-post imbalance across rounds. This is a concern because I add group member fixed effects and exploiting within-group member variation only. In other words, there are time effects as omitted variables that are large enough to reverse the results.

To address this concern, I re-estimate equation 2 with round fixed effects, and the results are presented in figure 8 in brown, labeled as "Time FE." The figure shows that this concern is invalid: all the coefficient estimates are about the same as the estimates of "Main spec." The confidence intervals are slightly wider because of the loss of degrees of freedom.

Significant non-linearity of correction? I included correction as a count variable in equation 2, so significant deviation from linearity can bias the results. To address this concern, I re-estimate equation 2 with correction as a dummy instead of count, and the results are presented in figure 8 in olive, labeled as "Dummy." The results show there is indeed non-linearity. First, the coefficient estimate on the female dummy is almost the same as those of "Main spec." Second, all correction terms – correction, correction interacted with the female dummy, and their sum –

are more negative. This suggests that the relationship between partner selection and correction term is concave – the first correction has the strongest effect than later corrections. Third, however, the coefficient estimate on the interaction between correction and female dummy is still statistically insignificant. Thus, although there is some non-linearity of correction, not taking into account for it simply makes the results conservative; we can get more quantitatively and statistically significant results when we incorporate the non-linearity.

Are unsolved puzzles driving the results? Because pairing is random, coefficients in equation 2 have a causal interpretation. However, the interpretation can be different if the results are driven by unsolved puzzles. This can mean two things: first, a correction occurs more often in unsolved puzzles and group members are less likely to select as a partner a participant with whom they could not solve the puzzle. Second, the contribution is not capturing actual contribution in that "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); indeed, group members are likely not to be able to solve the puzzle if the participant is not trying to solve the puzzle.

To address these concerns, I re-estimate equation 2 with solved puzzles only, and the results are presented in figure 8 in green, labeled as "Solved puzzles." The figure shows that these concerns are invalid: the correction effect gets stronger compared to the effect of "Main spec." However, the coefficient estimate on the interaction between correction and female dummy slightly gets positive albeit statistically and quantitatively insignificant. The confidence interval gets wider due to a drop in the number of observations by about 14%.

The figure also shows combinations of these alternative specifications of equation 2 but the results remain robust.

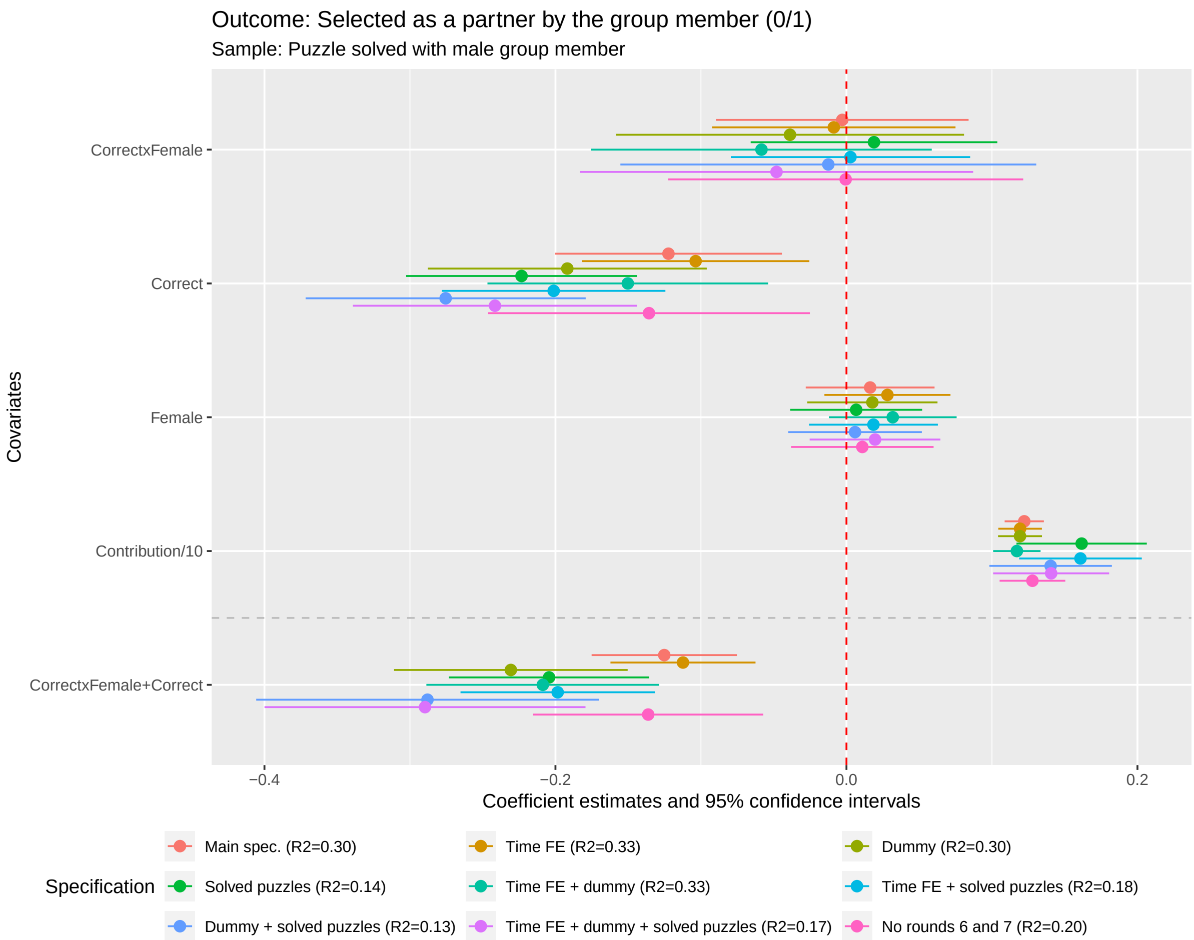
7.2 Robustness of the results for question 4

In figure 9, I present several alternative specifications of equation 4 to make sure that the results for question 4 I presented in column 3, table 6, are robust to alternative explanations. The y-axis presents covariates and the x-axis presents their coefficient estimates (points) and 95% confidence intervals (lines). To facilitate comparison, I again present results of column 3 of table 6 on the top in red labeled as "Main spec." Each specification address the same concerns as discussed in figure 8 in section 7.1; namely, (i) whether the negative correlation between correction and partner selection in rounds 6 and 7 are causal, (ii) whether ex-post imbalance across rounds is biasing the results, (iii) whether significant non-linearity in correction is biasing the results, and (iv) whether unsolved puzzles are driving the results. The results are robust to all these alternative explanations.

8 Discussion and conclusion

This paper studies women's cost of correcting male group members and its consequence on group efficiency. In order to study these questions, I design a quasi-laboratory experiment where participants are matched with seven other participants, solve one sliding puzzle together, and express a preference on which of them to be paired with in the final, payoff-relevant, part of the

Figure 9: Cost of correcting male group member's bad move (robustness)



Notes: This figure shows several alternative specifications of equation 4 to show the robustness of the results of column 3 of table 6 along with their adjusted R-squared. Specification "Main spec." is the column 3 results to make comparison easier. The y-axis shows covariates and the x-axis shows coefficient estimates of those covariates along with their 95% confidence intervals. The specification "Time FE" adds time fixed effects to equation 4 to show robustness to ex-post imbalance across rounds. The specification "Dummy" replaces correction in equation 4 with its dummy to show robustness to significant non-linearity. The specification "Solved puzzles" restricts the sample to solved puzzles only to show that the results are not driven by unsolved puzzles. The specification "No rounds 6 and 7" excludes observations in rounds 6 and 7 to show that the negative correlation between partner selection and correction in rounds 6 and 7 is not causation. Other specifications add combinations of these specifications. CR0 standard errors are clustered at the group member level.

experiment. I show that the paired participants' contribution to the puzzle is the most important factor for group members in selecting their partner and statistical discrimination story is unlikely: group members are equally likely to select women and men as a partner and there are no gender differences in contribution. Once I control for the paired participants' contribution to the puzzle, both male and female group members are significantly less likely to select a paired participant who corrected their move, regardless of the paired participants' gender. This reluctance to accept being corrected is efficiency reducing especially for male group members: male group members react more negatively to corrections that correct their wrong move. I show that the mechanism is male group members' overconfidence about their ability to solve the puzzle.

These results have three main implications for group work. First, correcting others should increase group efficiency in theory, but it is not necessarily so in the real world. Second, although women's corrections do not receive a stronger negative reaction, it may not be optimal for women to speak up more as the effect size of corrections is large. Third, men may be speaking up too much that reduces their and group's efficiency.

References

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

Ashford, Kate. n.d. "How to disagree with your boss without losing your job." *Monster.*

Ashraf, Nava, and Oriana Bandiera. 2018. "Social Incentives in Organizations." *Annual Review of Economics* 10 (1): 439–463.

Babcock, Linda, María P. Recalde, Lise Vesterlund, and Laurie Weingart. 2017. "Gender Differences in Accepting and Receiving Requests for Tasks with Low Promotability." *American Economic Review* 107 (3): 714–747.

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.

Boogaard, Kat. n.d. "How to Tell Your Boss "No"—Without Saying "No"." *The Muse.*

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

Born, Andreas, Eva Ranehill, and Anna Sandberg. 2020. "Gender and Willingness to Lead: Does the Gender Composition of Teams Matter?" *The Review of Economics and Statistics.*

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.

Carter, Alecia J., Alyssa Croft, Dieter Lukas, and Gillian M. Sandstrom. 2018. "Women's visibility in academic seminars: Women ask fewer questions than men." *PLOS ONE* 13 (9): e0202743.

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.

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

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

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

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.

Gallus, Jana, and Emma Heikensten. 2019. *Shine a light on the bright: The effect of awards on confidence to speak up in gender-typed knowledge work.* Working Paper.

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.

Greenwald, Morgan. 2018. "13 Clever Ways to Tell Your Boss "No"." *Best Life.*

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.

Haynes, Michelle C., and Madeline E. Heilman. 2013. "It Had to Be You (Not Me)!: Women's Attributional Rationalization of Their Contribution to Successful Joint Work Outcomes." *Personality and Social Psychology Bulletin* 39 (7): 956–969.

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.

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.

Lazear, Edward P., and Kathryn L. Shaw. 2007. "Personnel Economics: The Economist's View of Human Resources." *Journal of Economic Perspectives* 21 (4): 91–114.

Lebowitz, Shana, and Allana Akhtar. 2019. "How to say 'no' to your boss without looking lazy or incompetent." *Business Insider.*

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.

Marshall, Lisa B. 2020. "How to disagree with someone without seeming arrogant or angry." *Business Insider.*

McCord, Sara. n.d. “How to Disagree With Your Boss (Without Getting Fired).” *The Muse.*

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.

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.

Rosenberg McKay, Dawn. 2019. “How to Say No to Your Boss.” *The Balance Careers.*

Rudman, Laurie A., and Julie E. Phelan. 2008. “Backlash effects for disconfirming gender stereotypes in organizations.” *Research in Organizational Behavior* 28:61–79.

Sarsons, Heather, Klarita Gërxhani, Ernesto Reuben, and Arthur Schram. 2021. “Gender Differences in Recognition for Group Work.” *Journal of Political Economy* 129 (1): 101–147.

Shan, Xiaoyue. 2020. *Does Minority Status Drive Women Out Of Male-Dominated Fields?* 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.

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

Tajfel, Henry, and John Turner. 1979. “An integrative theory of intergroup conflict.” In *The social psychology of intergroup relations,* edited by William G. Austin and Stephen Worchel, 33–47. Monterey, CA: Brooks Cole Publishing.

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

Appendix A Definition of performance measures

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

$$\text{Player i's contribution} \equiv \frac{P_i}{P_i + P_j} \in [0,1],\ i,j = 1,2,\ i \neq j \tag{A1}$$

$$P_i \equiv \max\{\text{i's \# good moves} - \text{i's \# bad moves}, 0\}\ i = 1,2 \tag{A2}$$

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

Net good moves Net good moves are the number of good moves minus the number of bad moves a participant makes in a given puzzle in part 2. This is the expression inside the max operation in equation A2:

$$\text{Player i's net good moves} \equiv \text{i's \# good moves} - \text{i's \# bad moves} \in \mathbb{Z} \tag{A3}$$

Unconstrained contribution Unconstrained contribution is the same as contribution defined in equation A1, but not constrained between 0 and 1; in place of P_i and P_j, it uses net good moves for i and j in a given puzzle in part 2:

$$\text{Player i's unconstrained contribution} \equiv \frac{P_i'}{P_i' + P_j'} \in \mathbb{R},\ i,j = 1,2,\ i \neq j \tag{A4}$$

where P_i' is player i's net good moves. This measure is undefined when $P_i' + P_j' = 0$.

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 B 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).

Appendix C Further details of the ability measure selection

Table C1 presents results of running main regressions (equations 2 and 4) with ability measures other than contribution to elaborate ability measure selection in section 5.1: column 1 and 2 present results of equations 2 and 4 but with the number of puzzles solved alone as ability measure, columns 3 and 4 unconstrained contribution as ability measure, and columns 5 and 6 net good moves as ability measure.

Compared to no ability control reported in column 2 of Table 5, adding the number of puzzles solved alone only increases adjusted R-squared from 0.061 to 0.063 (column 1 of Table C1) and unconstrained contribution to 0.063 (column 3 of Table C1). Although R-squared is not a measure of goodness of fit for causal inference, because group members must care about participant's ability, a fair amount of variation of their partner selection must be explained by the participant's ability and must increase R-squared. These observations hold for another main regression (equation 4) reported in table 6.

On the other hand, adding net good moves increases adjusted R-squared from 0.061 to 0.304 (column 5 of Table C1), which is about the same increase compared to adding contribution which increases to 0.299 (column 3 of Table 5). However, net good moves have a long left tail and thus I consider it not a good ability measure to be controlled econometrically – it is difficult to give it a proper functional form. However, the results with net good moves as ability control give the same conclusions.

Table C1: Cost of correcting male group member's move (other ability measures)

Outcome:	Selected as a partner by the group member (0/1)					
Sample:	Puzzles solved with male group member					
Ability measure:	# puzzles solved in pt. 1		Contribution (unconstrained)		Net good moves	
	(1)	(2)	(3)	(4)	(5)	(6)
Correct×Female	-0.026	-0.110	-0.036	-0.136*	-0.021	0.053
	(0.045)	(0.083)	(0.044)	(0.082)	(0.035)	(0.071)
Correct	-0.153***	-0.271***	-0.148***	-0.257***	-0.070**	0.001
	(0.028)	(0.053)	(0.027)	(0.052)	(0.030)	(0.046)
Female	0.019	0.024	0.014	0.020	0.023	0.020
	(0.027)	(0.027)	(0.026)	(0.026)	(0.022)	(0.022)
CorrectGood×Female		0.084		0.109		-0.082
		(0.104)		(0.102)		(0.085)
CorrectGood		0.185***		0.171***		-0.104*
		(0.064)		(0.064)		(0.062)
Ability	0.014**	0.014**	0.124**	0.121*	0.078***	0.084***
	(0.007)	(0.007)	(0.063)	(0.062)	(0.003)	(0.003)
Group member FE	✓	✓	✓	✓	✓	✓
Correct×Female +Correct	-0.180***	-0.381***	-0.184***	-0.393***	-0.091***	0.054
	(0.036)	(0.061)	(0.035)	(0.061)	(0.024)	(0.058)
CorrectGood×Female +CorrectGood		0.270***		0.280***		-0.185***
		(0.079)		(0.077)		(0.067)
Baseline mean	0.747	0.747	0.747	0.747	0.747	0.747
Baseline SD	0.435	0.435	0.435	0.435	0.435	0.435
Adj. R-squared	0.063	0.077	0.063	0.077	0.304	0.308
Observations	1510	1510	1506	1506	1510	1510
Clusters	220	220	220	220	220	220

Notes: This table presents regression results of equations 2 and 4 but with ability measures other than contribution to show contribution is a good ability measure. Columns 1 and 2 present results of equations 2 and 4 but with the number of puzzles solved in part 1 as ability measure, columns 3 and 4 unconstrained contribution as ability measure, and columns 5 and 6 net good moves as ability measure. Baseline mean and standard deviation are that of men who do not correct group members. CR0 standard errors in parentheses are clustered at the group member level. Significance levels: * 10%, ** 5%, and *** 1%.

Gender differences in the cost of contradiction
Pre-analysis plan

Yuki Takahashi

November 22, 2020

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

1 Main Hypotheses

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

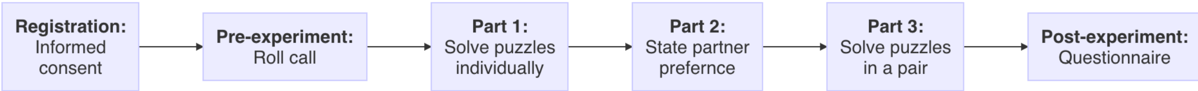
2 Design and procedure

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

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

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

Figure 1: Flowchart of the experiment



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

Registration

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

Pre-experiment

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

Part 1: Individual round

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

Part 2: Partner preference elicitation

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

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

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

   Part 3: Group round

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

   Post-experiment

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

3 Specification

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

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

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

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

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

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

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

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

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

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

other variables are as defined in equation 1.

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

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

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

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

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

other variables are as defined in equation 1.

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

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

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

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

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

References

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

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

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

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

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

Appendix A Power simulation

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

I assume the following data generating process:

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

where each variable is drawn from the following distribution:

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

Each parameter is defined as follows:

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

where L is total number of moves the decision-maker and a matched participant take to solve a puzzle, which I assume to be 15 (7.5 moves by the decision-maker). However, I also set it to 10 (5 moves by the decision-maker) for robustness check. $MDE = -0.02$ is my baseline assumption (being contradicted once reduces the probability of choosing a matched participant by the same degree as when the matched participant's contribution is 0.1 lower), but I also set it to -0.01 for robustness check, -0.03 to see what happens in a more optimistic scenario, and 0 to check that type I error rate is kept at 5% and that the estimated ATE is 0 when there is no underlying effect.

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

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

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

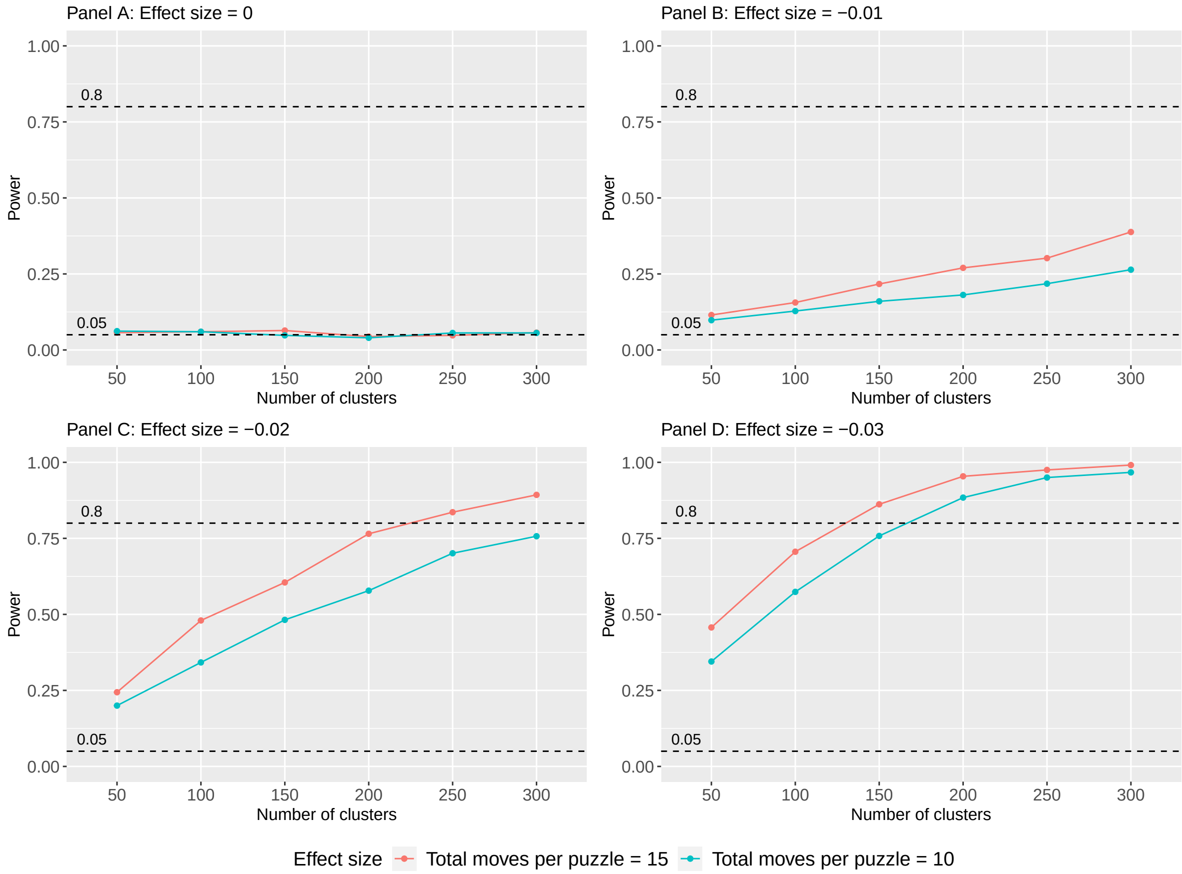
ATE is defined as the average of $\hat{\beta}_1$ across draws (its dependence on L is due to the non-linearity of the data generating process):

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

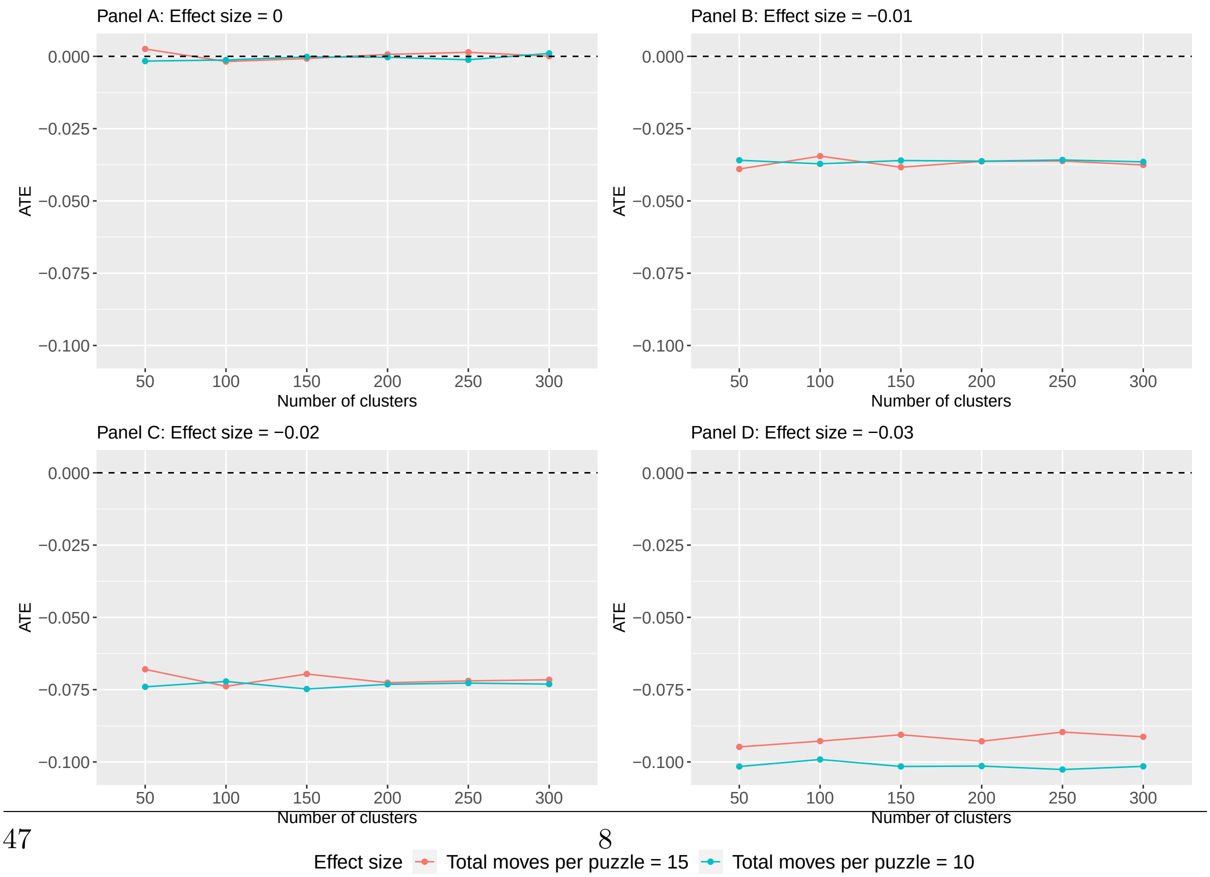
The results are presented in figure A1, which suggests that I need to recruit about 450 participants (so that I could have 225 clusters for testing H1). First, in the baseline scenario with $L = 15$, I can achieve about 80% power. Second, even under a tougher scenario where $L = 10$, I can still achieve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

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

(a) Estimated power



(b) Estimated ATE



Appendix B Questions asked in the questionnaire

English version

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

Italian translation

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

Appendix C Calculation of contribution

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

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

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

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