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 Please do not cite or circulate.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 4, 2021

Abstract

Women contribute their ideas less than equally knowledgeable men in a group. Since this reduces group efficiency and women's visibility in a group, behavioral interventions have been proposed to promote women's contribution. However, spelling out ones' ideas can result in correcting others, which in turn triggers others' negative feelings and hinders the effectiveness of the interventions. This paper studies women's cost of correcting male colleagues and its consequence to group efficiency in a setting where the correctness of corrections is only partially observable as in most group work. I design a quasi-laboratory experiment where participants first perform a joint task with several people who are their potential partner, select their partner from those people and work on the joint task with their partner. I show that the main determinant of participants' partner selection is a given person's contribution to the joint task and they correctly believe women and men are equally good at the joint task. However, *when comparing people with the same contribution*, participants are significantly less likely to select a person who has corrected their action as a partner regardless of that person's gender. Moreover, male participants react more negatively to a correction that fixes their mistake. The mechanism is participants' overconfidence about their ability. These findings suggest that behavioral interventions to increase women's contribution must be designed carefully as they may backfire, and that corrections, which should increase group efficiency in theory, do not necessarily do so in the real world.

JEL codes: J16, D91, C92
Keywords: correction, gender, 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, 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 contributing one's idea to a group is important for the group's and her or his own success: bringing together group members' expertise improves group decision making and group members recognize those who contribute to this process. Yet, women under-contribute their ideas relative to equivalently knowledgeable men (Coffman 2014); in academia, for example, women ask fewer questions than men during seminars (Carter et al. 2018; Dupas et al. 2021). This under-contribution is suboptimal for the group and the women themselves, and thus, studies propose behavioral interventions to increase their contribution (Coffman 2014; Gallus and Heikensten 2019).

However, a contribution of ideas can result in correcting others. In a setting where people cannot see whether the correction is right or wrong, as in most group work, correcting others can induce others' negative feelings. This possibility raises a concern that interventions to increase women's contribution may not be as effective as literature proposes both for the women and for the group: 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. This is a crucial problem because collaboration is necessary for career success; one's output depends on team members in group work and being selected into a team is essential to produce good outcomes. Even in academia where one's work is more individualistic, one can have more projects on their portfolio when their colleagues prefer them as a co-author which increases their chance to get tenured and to be more visible to other researchers.

This paper investigates 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 administrative 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. In the experiment, participants are matched in pairs and solve one sliding puzzle by alternating their moves. After solving the puzzle, participants decide whether they would like to be matched again with the person they were just matched with in a later stage which is the main source of earnings and thus gives a strong incentive for participants to select as a good collaborator as possible. Participants repeat this trial puzzle solving seven times, each time with a different participant in a random order, which addresses the first challenge.

I use Isaksson (2018)'s puzzle where participants solve a 3x3 sliding puzzle in pairs by alternating their moves and define a correction as reversing another participant's move, which addresses the second challenge. The puzzle also allows me to calculate an objective measure of each participant's contribution to the puzzle as well as to classify each move as good or bad, allowing me to keep individual contribution fixed and to determine whether a given correction improves group efficiency or not. Further, puzzle-solving captures an essential characteristic of group work that two or more people work together towards the same goal (Isaksson 2018) but the correctness of each move and corrections is only partially observable (but fully observable to the researcher). The outcome measure is being selected as a collaborator which comes directly after solving the puzzle and very relevant for one's career success as discussed above, which addresses the third challenge.

I find that the main determinant of participants' collaborator selection is matched participants' contribution to the puzzle, they believe women and men are equally good at the puzzle, and in fact, there are no gender differences in the contribution, suggesting that the collaborator selection is well-incentivized, participants partially observe other participants' ability through their moves, and that statistical discrimination story is unlikely. There is no gender difference in the propensity to correct other's moves either. However, after controlling for the matched participants' contribution, both male and female participants are less likely to select a matched participant who corrected their move as a collaborator 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, both male and female male participants react negatively to a correction regardless of the matched participants' gender.

This reluctance to accept being corrected may not reduce group efficiency if it is only about corrections that correct their good move. However, this is not the case; in fact, male participants react more negatively to a correction that corrects their wrong move than to a correction that corrects their right move. Female participants also react more negatively to such correction, but to a much lesser extent. Thus, while participants 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 collaborator and missing a successful teamwork opportunity.

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

Taken together, these findings suggest that behavioral interventions to increase women's contribution of ideas to a group can have a potential drawback, and must be tailored very carefully. Also, although corrections should increase group efficiency in theory, there are behavioral frictions that prevent achieving efficiency in the real world.

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 the 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 partner. However, on some occasions, the contribution of ideas has to be made openly, for example in academic seminars and business meetings, and in such cases, partners' response plays an important role in the efficiency of the interventions. 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 partner's wrong move in the same puzzle task I used in my experiment. My paper introduces correction in the contribution of ideas and examines its cost and effect on group efficiency.

More generally, my paper contributes to literature on gender differences in group work. Isaksson (2018) finds that women claim their contribution less 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 female economists when a paper is co-authored 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 introductory economics class when 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 enriches our understanding of gender differences in group work.

My paper also speaks to literature on social incentives in an organization (Ashraf and Bandiera 2018), in particular managerial favoritism. Literature develops theories that when objective measures of their performance are not available, 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), both of which reduce the efficiency of the organization. These claims are empirically verified in various settings (Bandiera, Barankay, and Rasul 2009; Beaman and Magruder 2012; Hjort 2014; Xu 2018). In addition, Li (2020) finds that favoritism not only distorts the optimal allocation of talent but also negatively affects non-favored workers' incentives. My paper suggests that one cause of such favoritism is a reluctance to accept being corrected.

2 Experiment

Introducing quasi-laboratory format I run the experiment in a quasi-laboratory format where we experimenters and the participants are connected 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.[1] On top of logistical convenience and complying with the COVID pre-caution measures, this quasi-laboratory format has the 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.

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

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 and 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). This is also a problem with any experiment 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 it takes time to collect a large amount of data.

Another benefit of quasi-laboratory experiments over standard online experiments is that we can screen participants based on their participation status in previous experiments which allows us to collect cleaner data; in particular, it 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 correctness of corrections objectively and comparably.[2] Further, puzzle-solving captures an essential characteristic of group work in which two or more people work towards the same goal (Isaksson 2018) but the correctness of each move and correction is only partially observable to participants (but fully observable to the experimenter).

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

Figure 1: Flowchart of the experiment

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

2.1 Design and procedure

Registration

Upon receiving an invitation email to the experiment, participants register for a session they want to participate in and upload their ID documents as well as a signed consent form.[3]

2. 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 s bad move is defined as a move that increases the number of moves away from the solution.

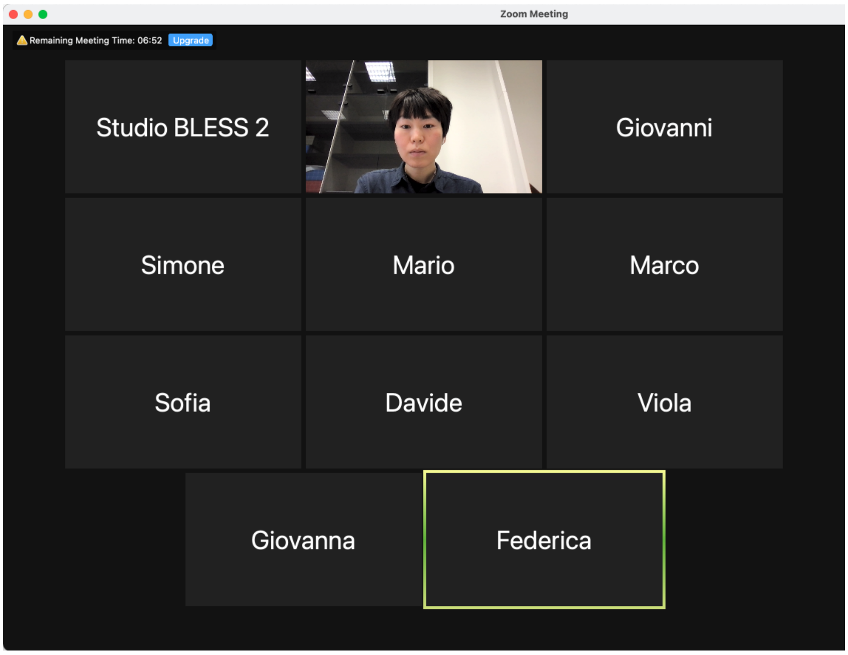
3. I recruit a few more participants than I need for a given session in case some participants would not show up to the session so that I could have 16 participants.

Pre-experiment

On the day and the time of the session of their choice, participants enter the Zoom waiting room.[4] 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.

After admitting all the participants to the Zoom meeting room, I do roll call (Bordalo et al. 2019; Coffman, Flikkema, and Shurchkov 2021): I 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 for the session, I draw random numbers from 1 to 40 and ask those who drew the coins with the same number to leave.[5] Those who leave the session receive the show-up fee. Figure 2 shows a Zoom screen participants would see during the roll call (the person with their camera 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 us only via Zoom's

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

5. I draw with replacement a number from 1 to 40; if no participant has a coin with the drawn number, I draw next number until the number of participants is 16.

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 collaborator

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

Second, participants are matched in pairs within each group and solve one puzzle together by alternating their moves. The participant to make 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 matched person's move.[6] 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 matched with another participant called Giovanni and waiting for Giovanni to make their move.

Once they finish the puzzle, participants state whether they would like to be matched again with the same person in part 3 (yes/no). At the end of the first round, new pairs are formed, with a perfect stranger matching procedure, do that every participant is matched 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 matched again with the same person 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.

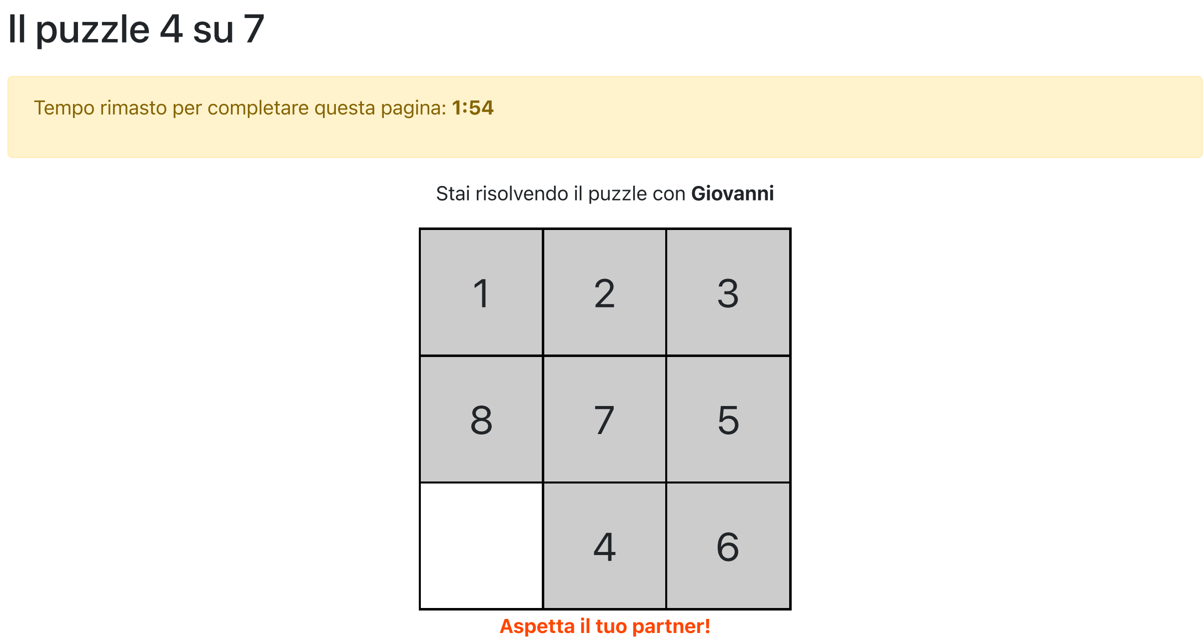
The matched person's first name is displayed on the computer screen throughout the puzzle and when participants select their collaborator to subtly inform the matched person's gender. Figure 4 shows an example of the collaborator selection screen where a participant finished playing a puzzle with another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

At the end of part 3, participants are matched 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 matched with i and with whom i is willing to be matched again in part 3.

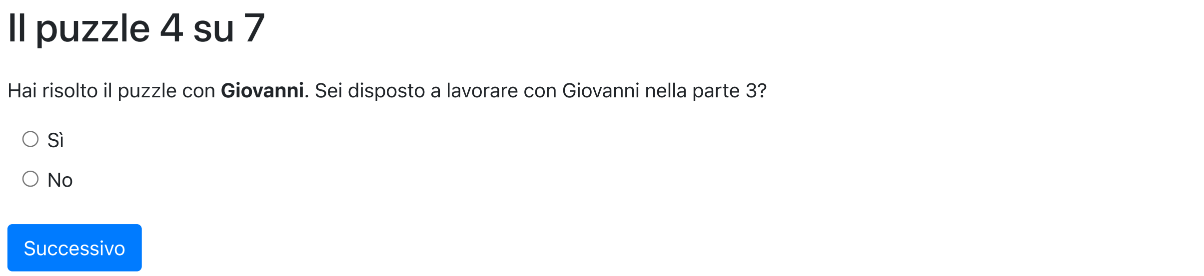
6. Solving the puzzle itself is not incentivized, and thus participants who do not want to work with the matched person or fear to receive a bad response may not reverse that person's move even if they think the move is wrong. However, since I am interested in the effect of correction on collaborator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

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 their move.

Figure 4: Collaborator selection screen



Notes: This figure shows a sample collaborator selection screen where a participant finished solving the 4th round puzzle with another participants called Giovanni and deciding whether to work with Giovanni 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 feasibly match is left.
6. If some participants are still left unmatched, I pair them up randomly.

Part 3: Solve puzzles with a collaborator

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) 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 the 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, which none has anticipated.

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).[7] The first two conditions were to reduce noise coming from differences in socio-demographic backgrounds and ethnicity that may be inferred from participants' first name and voice and the last condition was to reduce experimenter demand effects. The number of participants was determined by a power simulation in the pre-analysis plan to achieve 80% power.[8] The experiment and gender-related hypotheses are pre-registered with the OSF.[9]

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

I also use part 2 data only unless otherwise indicated. This is because it is part 2 where we can observe collaborator selection decisions.

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

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

8. 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.

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

3.2 Participants' characteristics

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.

Table 1 describes participants' characteristics. Male participants are slightly older than female participants by 1.4 years and more gender-biased. People from the south of Italy are slightly overrepresented for both female and male participants relative to the general Italian population.[10] 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).

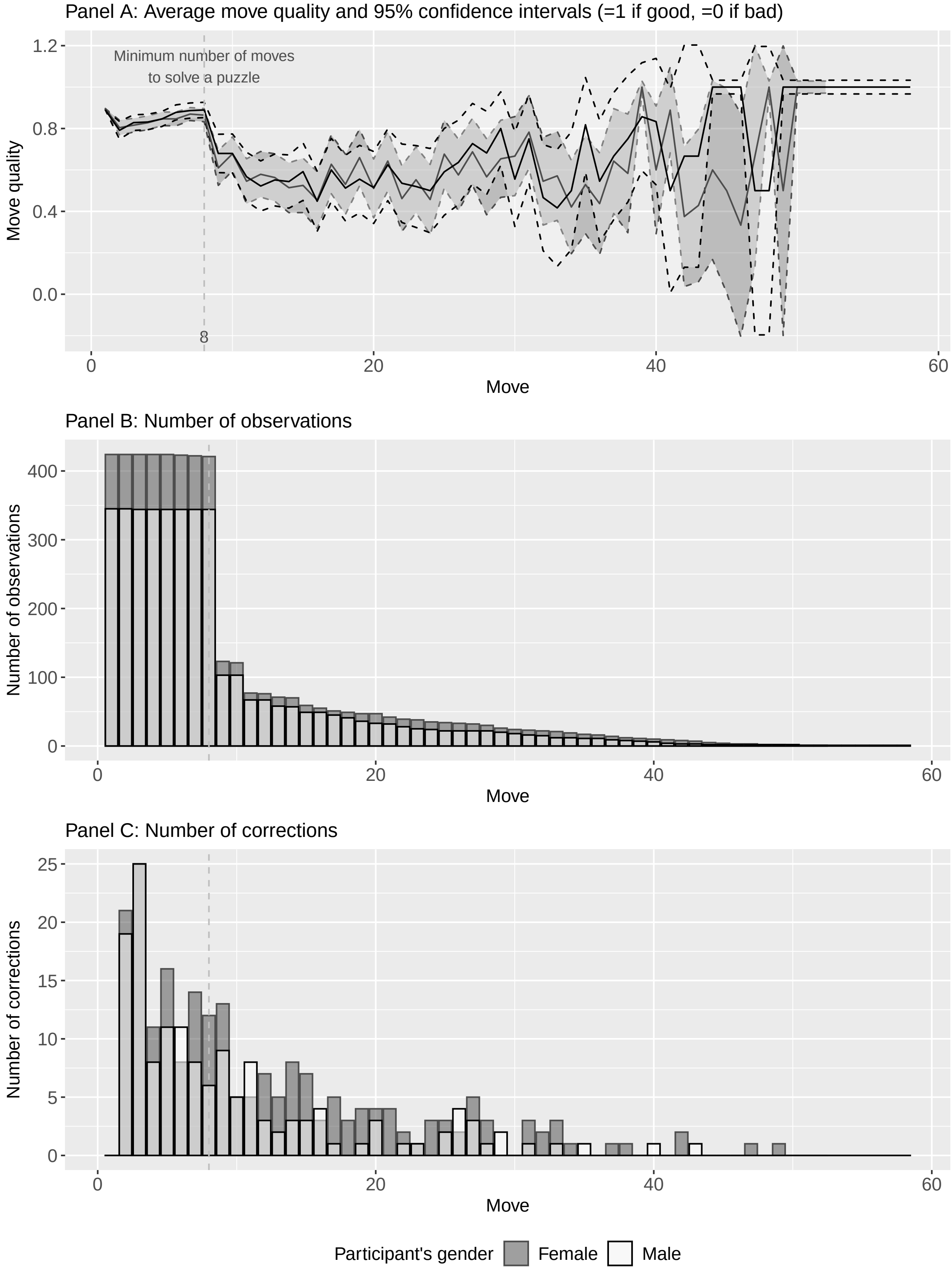
3.3 Move-level summary

Figure 5 shows average move quality along with 95% confidence intervals (panel A), number of observations (panel B), and number of corrections (panel C) for each move, separately for female (gray) and male participants (white). As mentioned before, I restrict the sample to puzzles where participants are matched with male participants.

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 gender

10. Despite that I recruited only Italy-born people, 1 male participant answered in the post-questionnaire that they were from abroad. I include this participant in the analysis anyway.

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 (panel B), and the number of corrections (panel C) for each move, separately for female (gray) and male participants (white). The confidence interval of panel A is 95% confidence intervals of *betas* 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 participants.

differences in the move quality. Panel B shows there are a little more than 400 moves up to the 8th move, then the number of moves 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, contribution, and collaborator selection.

3.4 Puzzle-level summary

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 alone	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: Partner'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 partner							
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 collaborator (0/1)	0.71	0.45	0.70	0.46	0.01	0.03	0.69
Selected as a collaborator (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 partner's puzzle-solving ability (panel B), corrections one made to partner (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 partner level. I restrict sample to puzzles where partners are male. Appendix A provides definitions of each puzzle-solving ability measure.

Table 2 describes own (panel A) and matched participants' puzzle-solving ability (panel B), corrections to partner (panel C), and puzzle outcomes (panel D). Panel A shows that female participants solve few 0.6 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 more later in figure 7.

Panel B shows that matched participants' (who are all male) puzzle performance in part 2 are not different for female and male participants. The difference in the number of puzzles solved in part 1 is statistically significant but is economically negligible.

Panel C shows that participants correct matched participants 15% of the times, of which 11-12% are good corrections (corrections of matched participants' wrong moves) and 4-5% are bad corrections (corrections of matched participants' right moves). Also, in some puzzles, participants make corrections more than once and those are mostly good corrections. There are no gender differences in the propensity of correction.[11]

Panel D shows that participants are selected as a collaborator by the matched participants 70-71% of the time. Even after netting out partner fixed effects,[12] the standard deviation of collaborator selection is high enough as shown in residualized collaborator selection, suggesting that adding partner fixed effects in the analysis does not control too much variation of collaborator selection. 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 matched participants correct each other's move consecutively. There is no gender difference in any of these puzzle outcomes.

3.5 Gender balance and puzzle outcomes across rounds

Remember that each participant plays the puzzle for seven rounds. 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.

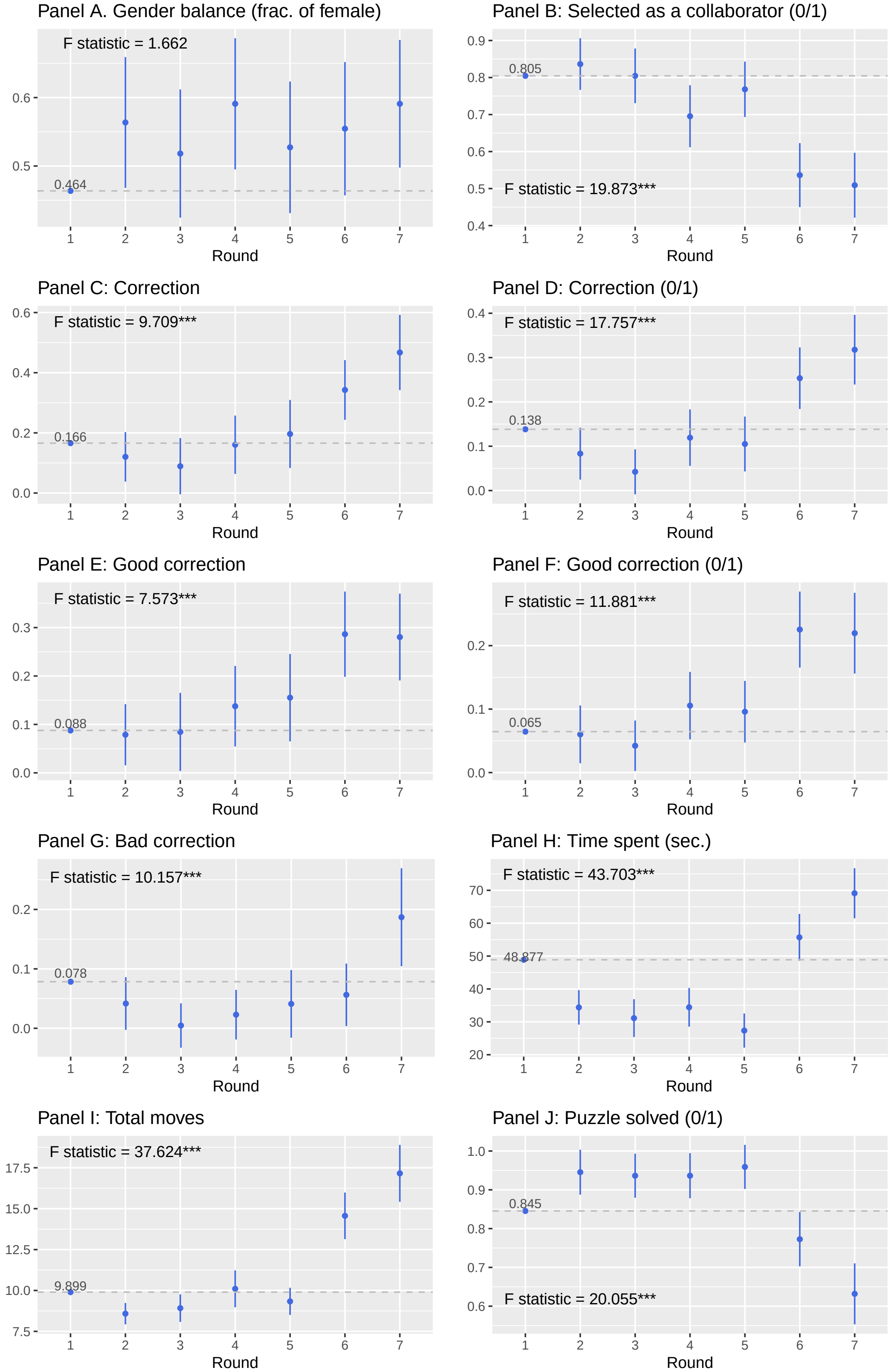
First, panel A shows that gender is roughly balanced across rounds as it should be because ex-ante everything is balanced by random matching. Second, panels B-J show that most outcome variables are unbalanced across rounds; specifically, whether a participant is selected as a collaborator and a puzzle is solved are both lower but in rounds 6 and 7, while the number of corrections, time a pair spend on the puzzle, and total moves – all of which are very likely to affect collaborator selection – are higher in rounds 6 and 7. It is unclear why there are these imbalances across rounds because all puzzles are the same difficulty: it could be that participants got tired in later rounds, puzzles in rounds 6 and 7 are perceived more difficult, etc.

However, the bottom line is that they are all outcomes of a particular match 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.

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

12. Residual from regressing collaborator selection on partner fixed effects.

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 number 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 indicates the joint significance of β_ks for 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 matched participant level. I restrict the sample to puzzles where participants are matched with male partner. Significance levels: * 10%, ** 5%, and *** 1%.

4 Theoretical framework

To yield questions to test and facilitate the interpretation of the results, I provide a simple theoretical framework.

I consider a participant i who maximizes their expected utility by selecting their collaborator j. i's utility depends on their payoff and emotional factor. The utility is increasing in the payoff and the payoff is increasing in i's belief about j's ability. Thus, if player i would select with whom to play in part 3, they would face the following problem:

$$\max_{j \in J} E_{\mu_i}[u_i(\pi(\mu_i(p_j, c_j, f_j)), \kappa_i(c_j, f_j))|\theta_i, \omega_i], \quad \partial u_i/\partial \pi > 0, \ \partial \pi/\partial \mu_i > 0 \tag{1}$$

where each term is defined as follows:

- $J \in \{1, 2, 3, 4, 5, 6, 7\}$: a set of all participants matched with i
- μ_i: i's belief about j's ability
- p_j: j's ability perceived by i
- c_j: j's correction (=1 if j corrected i, =0 not corrected)
- f_j: j's gender (=1 if female, =0 if male)
- θ_i: i's belief about their own ability relative to other participants (>0 if high, =0 if same, <0 if low)
- ω_i: i's belief about women's ability relative to men (>0 if high, =0 if same, <0 if low)
- κ_i: i's emotional factor

I make the following assumptions:

- μ_i is increasing in j's ability perceived by i: $\partial \mu_i/\partial p_j > 0$
- i's utility is decreasing in their emotional factor: $\partial u_i/\partial \kappa_i < 0$
- emotional factor is irrelevant if i is fully rational: $u_i(\pi, \kappa_i) \propto u_i(\pi)$

Because i can only partially observe j's ability, even if i is fully rational, j's correction, c_j, gender, f_j, and their interaction, $c_j * f_j$, also convey some information about j's ability.

4.1 When i is fully rational

Keeping j's perceived ability fixed, as I do in the analysis, the information j's correction conveys depends on i's belief about j's ability. If i believes they are good at the puzzle, they would consider correction as a signal of low ability because i believes their move is correct. On the other hand, if i do not believe that their ability is low, then they would consider correction as a signal of high ability. If they believe their ability is the same as others, then correction would not convey any information. Thus, $\partial \mu_i/\partial c_j < 0$ if $\theta_i > 0$, $\partial \mu_i/\partial c_j = 0$ if $\theta_i = 0$, and $\partial \mu_i/\partial c_j > 0$ if $\theta_i < 0$.

The information j's gender conveys depends on i's belief about women's ability to solve the puzzle relative to men. If i believes that women's ability is high, they would consider j being a woman as a signal of high ability. On the other hand, if i believes that women's ability is low, then they would consider j being a woman as a signal of low ability. If i believes that women's ability is the same as men's, then j being a woman would not convey any information. Thus, $\partial \mu_i/\partial f_j > 0$ if $\omega_i > 0$, $\partial \mu_i/\partial f_j = 0$ if $\omega_i = 0$, and $\partial \mu_i/\partial f_j < 0$ if $\omega_i < 0$. However, I control for j's gender in the analysis so j's gender itself does not matter in the analysis.

The information j's correction when j is a woman conveys depends on i's belief about their own ability and i's belief about women's ability. The role of both i's belief about their own ability and about women's ability is the same as correction in general and gender that I discussed above. Thus, $\partial^2\mu_i/\partial c_j\partial f_j > 0\ \forall\theta_i$ if $\omega_i > 0$, $\partial^2\mu_i/\partial c_j\partial f_j > 0\ \forall\theta_i$ if $\omega_i = 0$, and $\partial^2\mu_i/\partial c_j\partial f_j < 0\ \forall\theta_i$ if $\omega_i < 0$. These relationships hold for any θ_i because they compare women's correction relative to men's correction.

4.2 When i is not fully rational

When i is not fully rational, the emotional factor κ_i matters for their maximization problem. Specifically, I assume the following:

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

Both assumptions are based on literature on motivated reasoning. The first assumption is based on the finding that people consider those who disagree with them as biased (Kennedy and Pronin 2008). I assume that this belief affects i's actions. 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). I assume this belief affects i's actions. The second assumption can also be motivated by in-group/out-group bias (Tajfel and Turner 1979): when a person who belongs to an out-group (in this case women) does something bad to us, we react to it more negatively (similar to Chen and Li (2009)'s finding that people punish out-group members' misbehavior more).

4.3 Questions to test

To summarize, the theoretical framework yields the following questions (and expected direction for a fully rational participant):

Question 1. *Do men believe women and men are equally good at the puzzle? That is, is $\omega_i = 0$? — **Fully 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 collaborator a person who corrected their move? That is, is $\partial\mu_i/\partial c_j < 0$? — **Fully rational prediction**: Yes, men (and also women, to a lesser extent) are overconfident (Croson and Gneezy 2009) and thus likely to perceive correction as a signal of low ability.*

Question 3. *Are men less likely to select as a collaborator a woman than a man who corrected their move? That is, is $\partial^2\mu_i/\partial c_j\partial f_j < 0$? — **Fully rational prediction**: If the answer to question 1 is yes, then no.*

Remember that there are good and bad corrections and that participants can partially observe the quality of each move. Thus, although not modeled explicitly in equation 1, good corrections must convey more positive information about j 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? — **Fully rational prediction**: Yes, regardless of self-confidence and gender bias, participants should consider less negatively/more positively as a collaborator those who corrected their wrong move.*

5 Empirical strategy

Table 3: Conditions under which I need to observe partner's collaborator selection

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

Notes: This table shows conditions under which I need to observe partner's collaborator selection, keeping participant's perceived ability fixed. Participant's gender and correction must be exogenously varied. Partner must always be male.

Keeping participant's perceived ability fixed, I need to observe partner's collaborator selection in the four conditions shown in table 3 where participant's gender and correction are exogenously varied and the matched participant is always male. Comparing the conditions allows us to obtain the empirical counterparts of the relevant terms discussed in the theoretical framework:

- $C - D = \partial u_i / \partial f_j|_{c_j=0}$ (test of question 1)
- $(A + B) - (C + D) = \partial u_i / \partial c_j$ ($= \partial \kappa_i / \partial c_j$ if $\theta_i = 0$) (test of question 2)
- $(A - B) - (C - D) = \partial^2 u_i / \partial c_j \partial f_j$ ($= \partial^2 \kappa_i / \partial c_j \partial f_j$ if $\omega_i = 0$) (test of question 3)

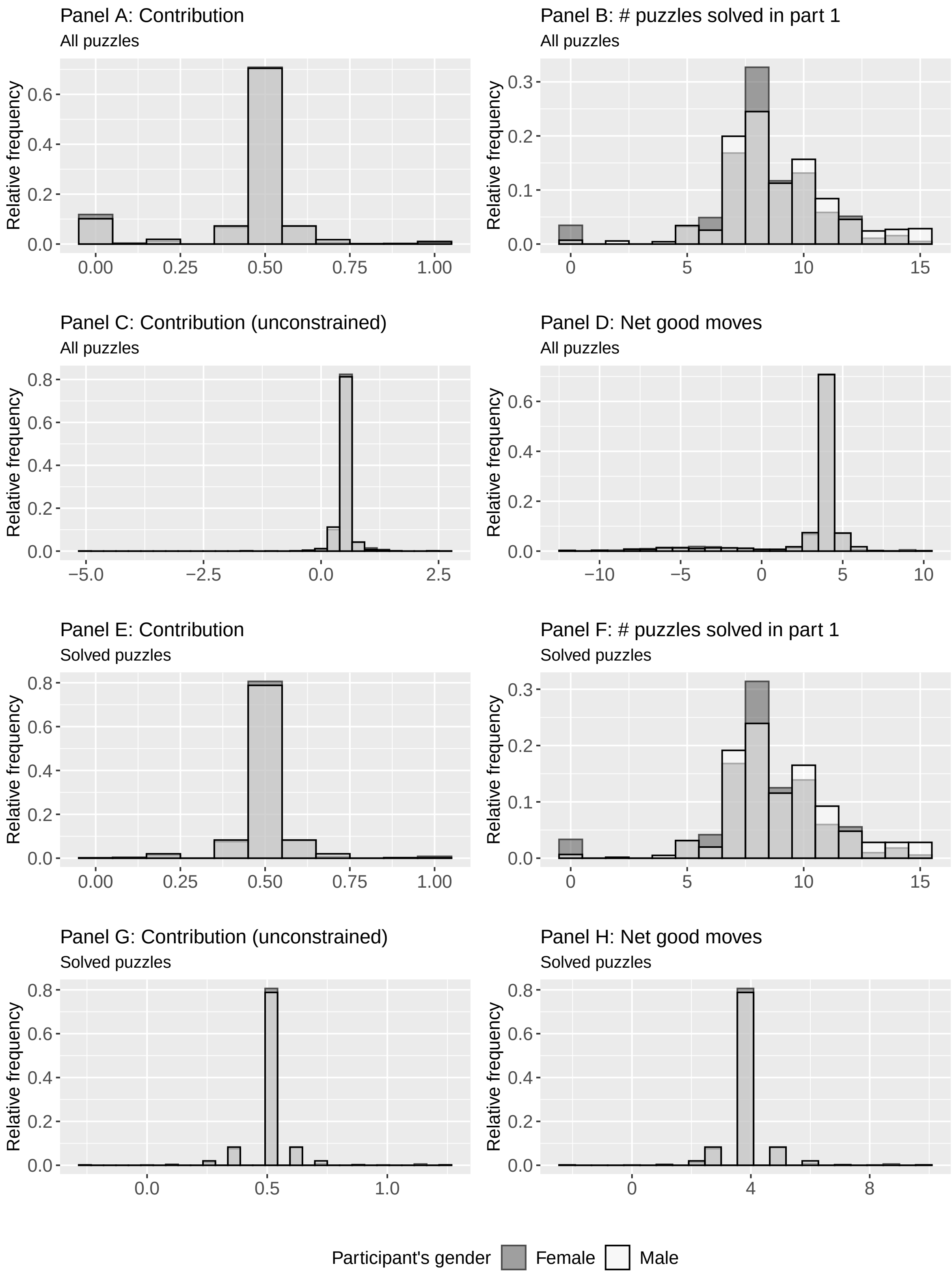
By random matching of participants, participant's gender is exogenous to matched participant's unobservables. However, correction is not exogenous. I could vary whether participants can correct matched participant, but it not only alters the rule of the puzzle (which alters participants' ability) but can also accumulate the matched participants' frustration for not being able to correct participants' move, which affects their collaborator selection but unobservable to me.

Thus, I instead allow participants to always correct the matched participants' moves and control for participant's ability both by design and econometrically. Because one's puzzle behavior affects the other's behavior and vice versa – for example, one's meanness can increase the other's correction and can also affect their collaborator selection – I include matched participant fixed effects. Under the assumption that participant's ability *perceived by the matched participant* is completely controlled for, correction happens by the participant's meanness, random move, perceived puzzle difficulty, etc., which are orthogonal to the matched participant's unobservables.

5.1 Ability measure selection

Figure 7 shows the empirical distributions of four ability measures: panels A-D include both solved and unsolved puzzles and panels E-H include solved puzzles only. While there are four

Figure 7: Gender differences in puzzle-solving ability



Notes: This figure shows 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.

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 well participant's ability in part 2. While unconstrained contribution and net good moves both capture participant's ability in part 2, they have a long left tail. In addition, matched participants will likely consider participants whose ability is really bad as equally unsuitable as a collaborator.

This point is more elaborated in table C1 in the appendix, where I run the main regression (equation 2) but with different ability measures. 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 matched participants must care participant's ability, a fair amount of variation of their collaborator 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. However, note that the results with net good moves as ability control give the same conclusions.

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 the most *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 matched with male participants: I only use female-male or male-male matches (so i is always male while j can be either female or male).

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

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

and $\mu_i \equiv \sum_{k=1}^{N} \mu^k \mathbb{1}[i = k]$ is i fixed effects, where N is the total number of i in the sample and $\mathbb{1}$ is the indicator variable. Standard errors are clustered at the matched participant level.[13]

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

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

Table 4: Conditions equation 2's coefficients identify

		Participant's gender Female	Male
Corrected partner (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 j's perceived ability fixed. Matched participants are always male.

Although I cannot estimate the intercept term, β_0, because of i fixed effects, I can still obtain the differences to test questions set up in section 4:

- $\beta_3 = \partial u_i/\partial f_j|_{c_j=0}$ (test of question 1)
- $\beta_1 + \beta_2 = \partial u_i/\partial c_j$ ($= \partial \kappa_i/\partial c_j$ if $\theta_i = 0$) (test of question 2)
- $\beta_1 = \partial^2 u_i/\partial c_j \partial f_j$ ($= \partial^2 \kappa_i/\partial c_j \partial f_j$ if $\omega_i = 0$) (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 bias from omitting some controls, column 4 shows the robustness of the column 3 results, and columns 5 and 6 show possible mechanisms. First, the most important thing to note is that contribution is the main determinant of male matched participant's decision to select a given participant, as the coefficient estimate on contribution is statistically and economically highly significant. This suggests that the collaborator selection is well-incentivized and that participants can partially observe other 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 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 matched participant fixed effects seem to make correction effect weaker.

Third, column 4 adds the number of puzzles solved in part 1 as an additional puzzle ability control to see whether part 1 behavior affects matched participants' collaborator selection. However, compared to column 3, adding the number of puzzles solved in part 1 does not change almost anything: 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 matched participant's collaborator selection.

Below, I discuss how column 3 answers each of the questions I posed in section 4 below. I also discuss what columns 5 and 6 tell us about the mechanism.

Table 5: Cost of correcting male partner's move

Outcome:	Selected as a collaborator (0/1)						
Sample: Puzzles solved with...	male partner						female and male partner
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
CorrectxFemale	-0.017	-0.027	-0.003	-0.003	0.006		
	(0.044)	(0.045)	(0.044)	(0.044)	(0.069)		
Correct	-0.149***	-0.152***	-0.122***	-0.123***	-0.131**	-0.179***	-0.125***
	(0.025)	(0.028)	(0.040)	(0.040)	(0.061)	(0.029)	(0.024)
Female	0.017	0.013	0.016	0.020	0.003	0.014	0.009
	(0.024)	(0.026)	(0.022)	(0.023)	(0.034)	(0.021)	(0.014)
Contribution			1.222***	1.218***	1.221***	1.230***	1.199***
			(0.068)	(0.068)	(0.069)	(0.068)	(0.050)
# puzzles solved in pt. 1				0.008			
				(0.006)			
PtrHighBiasxCorrectxFemale					-0.013		
					(0.089)		
PtrHighBiasxCorrect					0.013		
					(0.080)		
PtrHighBiasxFemale					0.019		
					(0.045)		
PtrLowAbilityxCorrect						0.083*	
						(0.042)	
PtrFemalexCorrect							-0.067**
							(0.030)
Partner FE	-	✓	✓	✓	✓	✓	✓
CorrectxFemale+Correct	-0.166***	-0.179***	-0.125***	-0.126***	-0.125***		
	(0.035)	(0.036)	(0.025)	(0.025)	(0.047)		
PtrHighBiasxCorrectxFemale +CorrectxFemale					-0.007		
					(0.057)		
PtrLowAbilityxCorrect +Correct						-0.096***	
						(0.031)	
PtrFemalexCorrect +Correct							-0.191***
							(0.019)
Baseline mean	0.747	0.747	0.747	0.747	0.749	0.747	0.745
Baseline SD	0.435	0.435	0.435	0.435	0.434	0.435	0.436
Adj. R-squared	0.047	0.061	0.299	0.300	0.298	0.302	0.318
Observations	1510	1510	1510	1510	1503	1510	3180
Clusters	220	220	220	220	219	220	464

Notes: This table presents regression results of equation 2. Column 1 does not control for contribution to the puzzle and does not have matched participant fixed effects, column 2 controls for contribution to the puzzle but does not have matched participant fixed effects, and column 3 controls for contribution to the puzzle and has matched participant 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 matched participants with higher and lower gender bias to see whether male matched participants with higher bias respond more negatively/less positively to women's corrections. Column 6 separates the coefficient estimates for male matched participants with higher and lower ability to see whether the results in column 3 are driven by their overconfidence; interaction between correction and female dummy is dropped to increase efficiency and not to pick up an imbalance in the data. Column 7 includes both female and male matched participants (that is, $i \in \{Female, Male\}$) and examines gender differences in response to correction. Baseline mean and standard deviation are that of men who do not correct matched participants. CR0 standard errors in parentheses are clustered at the partner level. Significance levels: * 10%, ** 5%, and *** 1%.

6.1 Do men believe women and men are equally good at the puzzle? – Formal test of 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 matched participants select women over men as their collaborator without correction effect. As we see in column 3, it is statistically and economically insignificant; in fact, it is slightly positive. Thus, absence of correction, men are not less likely to select women over men as their collaborator. Because there is no reputation concern, there is no reason for a matched participant to select a participant whom they believe has a lower ability. Thus, the results suggest that men believe women and men as equally good at the puzzle.

As we see in table 2 and figure 7, there are no gender difference in puzzle ability. Although men solved slightly more puzzles in part 1, participants do not observe other participants' part 1 behavior and if we restrict our attention to part 2 ability, there is no gender differences both in mean and in distribution. Because participants can partially observe other participants' abilities, this evidence further backs up the claim that men believe women and men as equally good at the puzzle. Thus, statistical discrimination story is unlikely to be the explanation of any gender differences.

6.2 Are men less likely to select as a collaborator 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 participant's correction on a male matched participant's probability to select that participant as a collaborator, regardless of that participant's gender. It is negative and statistically and economically 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 matched male participant's probability to select a man who does not correct them), the effect is 16.3%.

Remember that this effect is causal, that it is comparing the two participants who contributed the same to the puzzle but one corrected the matched participant 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: $|\hat{\beta}_2|/(\hat{\delta} * SD_{contribution,male}) = |-0.122|/(1.222 * 0.17) \approx 0.59$. So one has to increase their contribution by 0.59 standard deviation to be equally preferred by the matched participant as a collaborator, which is pretty large.

6.3 Are men less likely to select as a collaborator 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 matched participant's probability of select her as a collaborator. Although it is negative, it is neither statistically nor economically significant. This suggests that women's correction does not receive a stronger negative reaction relative to men's.

Nevertheless, 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 economically highly significant. Thus, even if women's correction does not necessarily receive a stronger reaction, this result suggests that behavioral interventions to promote women's contribution must be tailored very carefully.

Heterogeneity by men's gender bias Matched male participants 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 matched participants with higher and weaker gender bias, which I do by augmenting equation 2 and re-estimating it via OLS:

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

where each variable is defined as follows:

- $HighBias_i \in \{0, 1\}$: an indicator variable equals 1 if i'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 PrtHighBias show the difference of the effect of those terms for matched male participants with higher and lower gender bias. In short, I do not find any statistically or economically significant difference for any effect. This could be due to that the gender bias questions honestly because it is a socially sensitive issue.

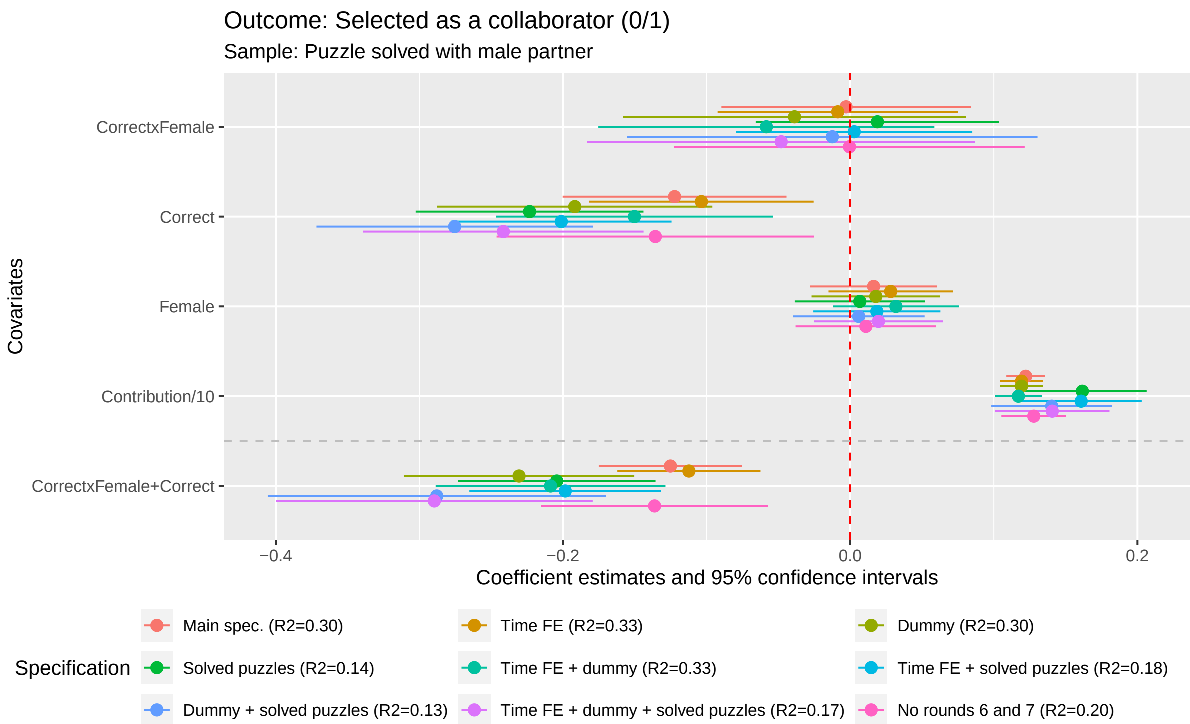
6.4 Robustness of the results for questions 1-3

I present several alternative specifications of equation 2 to make sure that the results for questions 1-3 I presented so far are robust to alternative explanations in figure 8. The y-axis presents covariates and the x-axis presents their coefficient estimates (dots) 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 collaborator 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 controlling for the observed ability enough 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

Figure 8: Cost of correcting male partner'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 collaborator 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 partner level.

that this concern is invalid: all the coefficient estimates are about the same as the estimates in column 3 of table 5 which are presented again in this figure as "Main spec." for comparison. The confidence intervals are wider due to a drop in the number of observations by 2/7.

Ex-post imbalance across rounds? Although ex-ante all rounds must be balanced by random matching, one may concern that there is an ex-post imbalance across rounds. This is a concern because I add matched participant fixed effects and exploiting within-matched participant 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 in column 3 of table 5 which are presented again in this figure as "Main spec." for comparison. 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 non-linearity. First, the coefficient estimate on the female dummy is almost the same as column 3 of table 5 which is presented again in this figure as "Main spec." for comparison. Second, all correction terms – correction, correction interacted with the female dummy, and their sum – are more negative. This suggests that the relationship between collaborator selection and correction term is concave – the first correction has a stronger effect than the second and later corrections. Third, however, coefficient estimate on the interaction between correction and female dummy is still statistically insignificant. Thus, there is some non-linearity of correction, but it just makes the results more conservative; we can get more economically and statistically significant results when we incorporate the non-linearity.

Are unsolved puzzles driving the results? Because matching 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 matched participants are less likely to select as a collaborator 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).

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 in column 3 of table 5 which is presented again in this figure as "Main spec." for comparison. However, the coefficient estimate on the interaction between correction and female dummy slightly gets positive albeit statistically and economically 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.

6.5 Does a correction that corrected 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 collaborator 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:

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

Table 6: Cost of correcting male partner's bad move

Outcome:	Selected as a collaborator (0/1)					
Sample: Puzzles solved with...	male partner					female and male partner
	(1)	(2)	(3)	(4)	(5)	(6)
CorrectxFemale	-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	-0.054	-0.028
	(0.051)	(0.052)	(0.046)	(0.047)	(0.034)	(0.034)
Female	0.023	0.018	0.013	0.016	0.014	0.009
	(0.024)	(0.026)	(0.022)	(0.023)	(0.021)	(0.014)
CorrectGoodxFemale	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**	-0.182***	-0.136***
	(0.064)	(0.064)	(0.061)	(0.062)	(0.054)	(0.044)
Contribution			1.304***	1.300***	1.312***	1.246***
			(0.077)	(0.077)	(0.076)	(0.054)
# puzzles solved in pt. 1				0.007		
				(0.006)		
PtrLowAbilityxCorrect					0.062	
					(0.057)	
PtrLowAbilityxCorrectGood					0.041	
					(0.077)	
PtrFemalexCorrect						-0.106**
						(0.046)
PtrFemalexCorrectGood						0.060
						(0.059)
Partner FE	-	✓	✓	✓	✓	✓
CorrectxFemale+Correct	-0.392***	-0.382***	0.004	0.004		
	(0.055)	(0.061)	(0.055)	(0.055)		
CorrectGoodxFemale+CorrectGood	0.291***	0.273***	-0.169***	-0.170***		
	(0.073)	(0.078)	(0.066)	(0.065)		
PtrLowAbilityxCorrect					0.008	
+Correct					(0.052)	
PtrLowAbilityxCorrectGood					-0.141**	
+CorrectGood					(0.064)	
PtrFemalexCorrect						-0.134***
+Correct						(0.036)
PtrFemalexCorrectGood						-0.076*
+CorrectGood						(0.044)
Baseline mean	0.747	0.747	0.747	0.747	0.747	0.745
Baseline SD	0.435	0.435	0.435	0.435	0.435	0.436
Adj. R-squared	0.064	0.075	0.304	0.304	0.307	0.321
Observations	1510	1510	1510	1510	1510	3180
Clusters	220	220	220	220	220	464

Notes: This table presents regression results of equation 4. Column 1 does not control for contribution to the puzzle and does not have matched participant fixed effects, column 2 controls for contribution to the puzzle but does not have matched participant fixed effects, and column 3 controls for contribution to the puzzle and has matched participant 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 matched participants with higher and lower ability to see whether the results in column 3 are driven by their overconfidence; the interaction between good and bad correction and female dummy is dropped to increase efficiency and not to pick up an imbalance in the data. Column 6 includes both female and male matched participants (that is, $i \in \{Female, Male\}$) and examines gender differences in response to correction. Baseline mean and standard deviation are that of men who do not correct matched participants. CR0 standard errors in parentheses are clustered at the partner level. Significance levels: * 10%, ** 5%, and *** 1%.

where each variable is defined as follows:

- $CorrectGood_{ij} \in \{0, 1, ...\}$: the number of times j corrects i'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 matched participant 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 matched participants' collaborator selection. Column 5 shows a possible mechanism, which I will explain later.

The coefficient estimate on correction of column 3 of table 6 (β_2 in equation 4) is the effect of a participant's good correction on the male matched participant's probability to select that participant as a collaborator relative to a bad correction, regardless of that participant's gender. It is negative, economically 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 the same to the puzzle but one corrected the matched participant's bad move and the other corrected 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 into teamwork and bad correction decreases the probability.

Thus, while participants 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 collaborator 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.

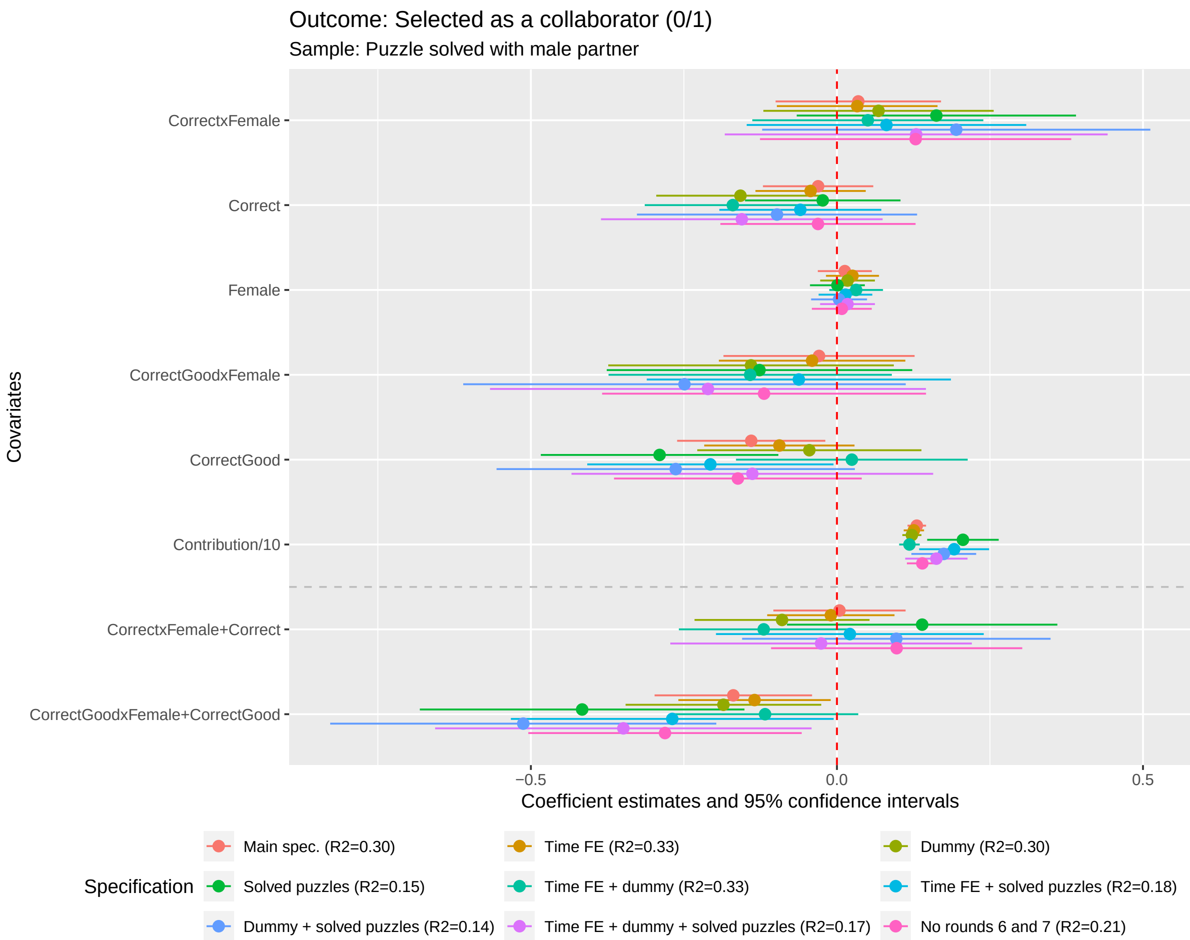
6.6 Robustness of the results for question 4

I present several alternative specifications of equation 4 in figure 9 to make sure that the results for question 4 I presented so far are robust to alternative explanations. The y-axis presents covariates and the x-axis presents their coefficient estimates (dots) 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." Each specification address the same concerns as discussed in figure 8 in section 6.4; namely, (i) whether the negative correlation between correction and collaborator selection in rounds 6 and 7 are causal, (ii) whether ex-post imbalance across rounds is biasing the results, (iii) significant non-linearity in correction, and (iv) whether unsolved puzzles are driving the results. The results are robust to all these alternative explanations.

6.7 Why do men react negatively to corrections?

Until now, we see that it is efficiency reducing for male matched participants to be reluctant to accept being corrected. One remaining question is why. One possible reason is overconfidence:

Figure 9: Cost of correcting male partner'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 collaborator 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 partner level.

remember that the correctness of the correction is not fully observable and corrections convey information about the participant's ability as we saw in the theoretical framework in section 4. Also, remember that participants are told how many puzzles they have solved in part 1 at the end of that part. Thus, those who solved a fewer number of puzzles in part 1 should be less confident about their ability than those who solved a larger number of puzzles. This does not affect the other participants' ability relative to them because participants are grouped with other participants of similar abilities. Thus, if their reluctance is driven by their overconfidence, those who solved a fewer number of puzzles should respond less negatively to corrections than those who solved a larger number of puzzles because they are less likely to believe their moves to be correct.

I test this conjecture by separating the effect of correction for matched male participants who solved fewer and larger number of puzzles, which I do 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 with data with a moderate number of clusters such as mine.

$$Selected_{ij} = \beta_0 + \beta_1 Correct_{ij} + \beta_2 Female_j + \beta_3 Correct_{ij} * LowAbility_i + \delta Contribute_{ij} + \mu_i + \epsilon_{ij} \\ i \in \{Male\}, j \in \{Female, Male\} \tag{5}$$

where each variable is defined as follows:

- $LowAbility_i \in \{0, 1\}$: an indicator variable equals 1 if i is grouped with low ability participants, 0 otherwise.

other variables are defined as in equation 2.

The results are presented in column 6 of table 5. First, the coefficient estimate on correction, which is the effect of correction on high ability matched participants, is more negative than the estimate in column 3. Second, the coefficient estimate on the interaction between correction and low ability matched participants is positive and statistically significant at 10%. These results suggest that it is male matched participants' overconfidence that is driving their reluctance to accept being corrected.

However, because low ability matched participants are low ability – they cannot observe move quality as well as high ability matched participants – 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 matched participants' negative reaction to bad correction.

To test this possibility, I separate the effect of good and bad correction for high and low ability matched participants. If the above finding comes from ability difference, then high ability matched participants 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.

$$\begin{aligned} Selected_{ij} =& \beta_0 + \beta_1 Correct_{ij} + \beta_2 Female_j + \beta_3 CorrectGood_{ij} + \beta_4 Correct_{ij} * LowAbility_i \\ &+ \beta_5 CorrectGood_{ij} * LowAbility_i + \delta Contribute_{ij} + \mu_i + \epsilon_{ij} \\ & i \in \{Male\}, j \in \{Female, Male\} \end{aligned} \tag{6}$$

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

The results are presented in column 5 of table 6. First, the coefficient estimate on correction, which is the effect of bad correction on high ability matched participants, is slightly more negative than the estimate in column 3. Second, the coefficient estimate on good correction, which is the difference between the effect of good and bad correction on high ability matched participants, is more negative than the estimate in column 3. Third, the coefficient estimate on the interaction between correction and low ability matched participants, which is the difference between the effect of bad correction on high and low ability matched participants, is positive although statistically insignificant. Fourth, the coefficient estimate on the interaction between good correction and low ability matched participants, which is a double difference between the effect of good and bad correction on high and low ability matched participants, is positive although statistically insignificant. These are *inconsistent* with the story that the results we saw in column 6 of table 5 are simply due to that high ability matched participants can observe move quality better than low ability matched participants; if so, high ability matched participants must respond less negatively to good corrections.

6.8 Gender differences in responses to corrections (exploratory)

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

Response to correction To test gender differences in corrections, I pool both female and male matched participants and separate the effect of correction for matched female and male participants. 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 make interpretation easier.

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

where each variable is defined as follows:

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

other variables are defined as in equation 2.

Column 7 of table 5 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. Second, however, the coefficient estimate on the interaction between correction and female matched participant dummy is negative and statistically significant at 5%, suggesting that women react more negatively to correction than men. It is economically significant too: correcting a female group member reduces the probability of being selected into a teamwork by 19.1 percentage points. Relative to the baseline mean (the matched male participant's probability to select a man who does not correct them), the effect is 25.6%.

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

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

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

Column 6 of table 6 presents the results of equation 8. First, the coefficient estimate on the interaction between correction and female matched participant 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 matched participant dummy is positive and economically significant, although statistically insignificant. Third, looking at the sum of coefficient estimates on the interaction between good correction and female matched participant dummy and on good correction, it is negative but less so than male matched participants, and statistically significant only at 10%.

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.

6.9 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 stake is much smaller: it is just a puzzle after all and not something people have been devoting much of their time, for example, research projects and corporate investment projects.

7 Discussion and conclusion

This paper has studied 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 select their team member with whom they solve more puzzles together from among those matched participants. I show that the matched participants' contribution to the puzzle as the most important factor for participants in selecting their team member, they believe women and men are equally good at the puzzle, and in fact, there are no gender differences in contribution. Once I control for the matched participants' contribution to the puzzle, both male and female participants are significantly less likely to select a matched participant who corrected their move, regardless of the matched participants' gender. This reluctance to accept being corrected is efficiency reducing: male participants react more negatively to corrections that correct their wrong move. I show that the mechanism is male participants' overconfidence about their ability to solve the puzzle.

These results have two main implications for group work. First, while literature finds that women under contribute their ideas in group work and suggests behavioral interventions to promote it, it has to be done very carefully. One way of such intervention is Gallus and Heikensten (2019) where they put women's ideas ahead of men's without openly correcting men. Second, correcting others should increase group efficiency in theory, but it is not necessarily so in the real world.

References

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

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.

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.

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.

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.

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.

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.

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:

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

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

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:

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

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:

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

This measure is undefined when the denominator is 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 US undergraduate students' degree of sexism.

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. I use a certified Italian translation from Manganelli Rattazzi, Volpato, and Canova (2008) and Rollero, Glick, and Tartaglia (2014) in the experiment.

> 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.

Appendix C Results with other ability measures

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

Outcome:	Selected as a collaborator (0/1)					
Sample:	Puzzles solved with male partner					
Ability measure:	# puzzles solved in pt. 1		Contribution (unconstrained)		Net good moves	
	(1)	(2)	(3)	(4)	(5)	(6)
CorrectxFemale	-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)
CorrectGoodxFemale		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)
Partner FE	✓	✓	✓	✓	✓	✓
CorrectxFemale+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)
CorrectGoodxFemale+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 equation 4 but with ability measures other than contribution to show contribution is a good ability measure. Columns 1 and 2 present results of equation 2 and 4 but with number of puzzles solved alone as ability measure, column 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 matched participants. CR0 standard errors in parentheses are clustered at the partner level. Significance levels: * 10%, ** 5%, and *** 1%.

Table C1 presents results of running main regression equations 2 and 4 with ability measures other than contribution: 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.

As discussed in section 5, the number of puzzles solved alone and unconstrained contribution increase adjusted R-squared only very little compared to no ability control reported in column 2 of table 5 for equation 2 and column 2 of table 6 for equation 4. R-squared is not a measure of goodness of fit for causal inference, but matched participants must care participant's ability, so a fair amount of variation of their collaborator selection must be explained by participant's ability and must increase R-squared. Hence, I consider these ability measures not appropriate.

Net good moves increase adjusted R-squared as much as contribution, but it has a long left tail issue. Still, the results with net good moves as ability control give the same conclusions as the main results.

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.

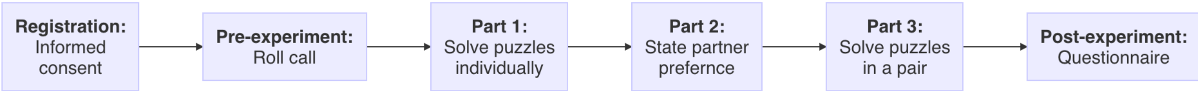
2 Design and procedure

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:

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

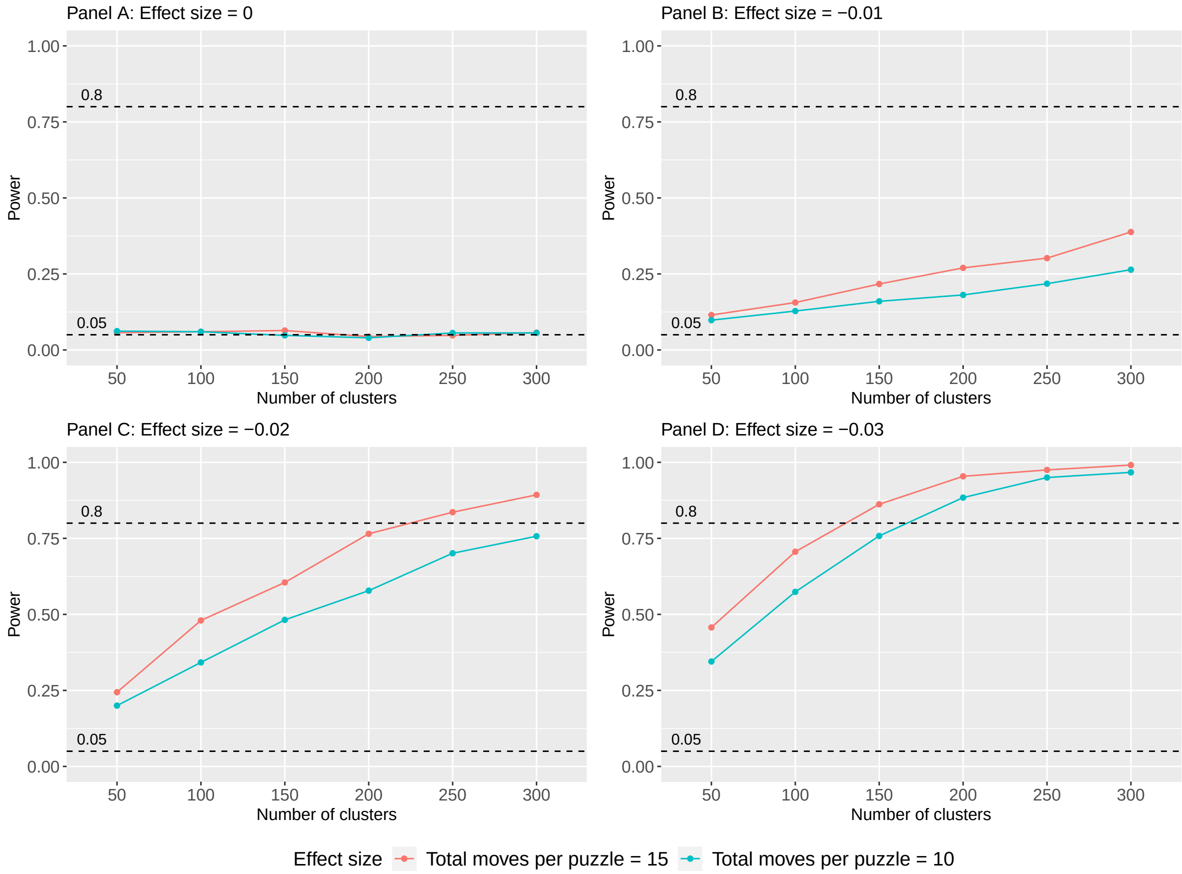
ATE is defined as the average of $\hat{\beta}_1$ across draws (its dependence on L is due to the non-linearity of the data generating process):

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

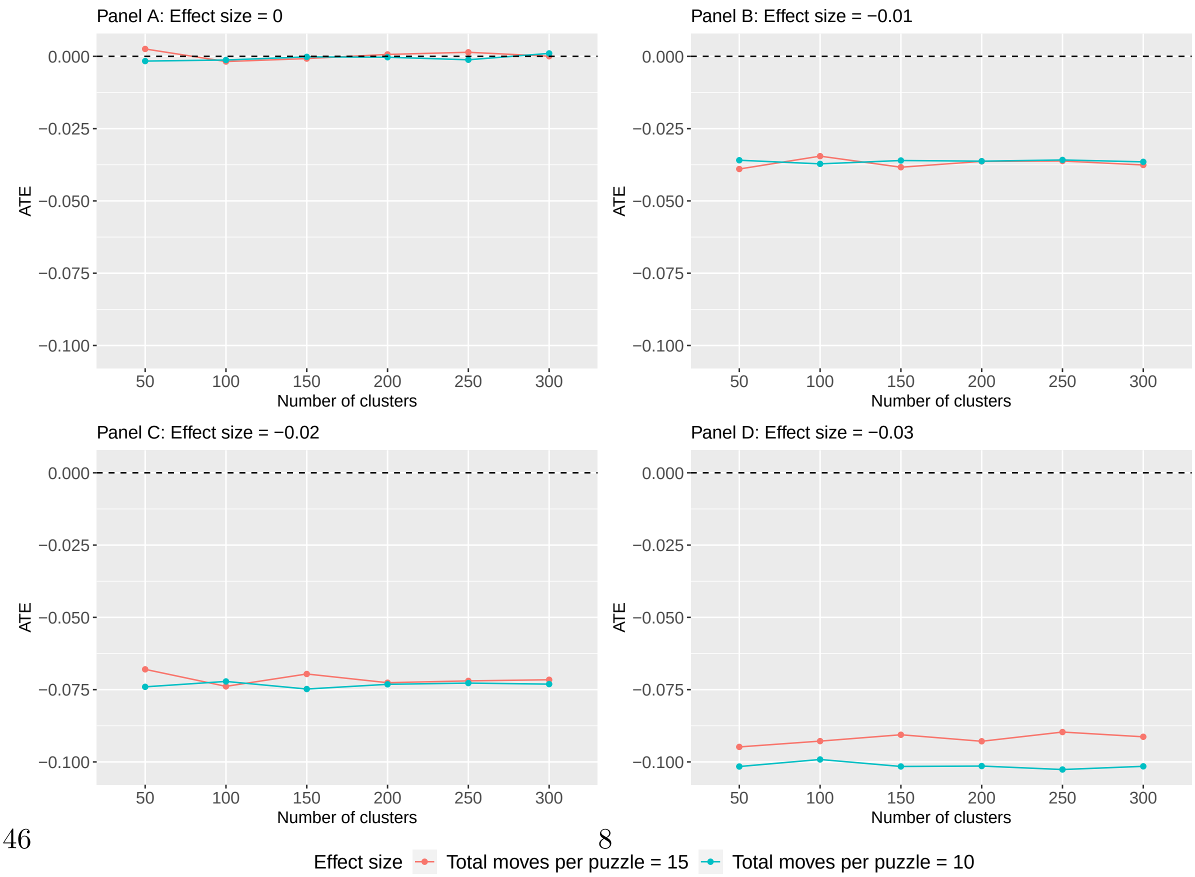
The results are presented in figure A1, which suggests that I need to recruit about 450 participants (so that I could have 225 clusters for testing H1). First, in the baseline scenario with $L = 15$, I can achieve about 80% power. Second, even under a tougher scenario where $L = 10$, I can still achieve 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.