The Cost of Corrections in Group Work

Yuki Takahashi*

Preliminary draft. Please do not cite or circulate.

Click here for the latest version

June 4, 2021

Abstract

Collecting colleagues could improve group efficiency. However, it may make them emotionally irritated and reduce the chance to be selected into teamwork. This paper studies how being corrected by others in a group affects one's probability of selecting that person as a partner in later works. I design a quasi-laboratory experiment where people first perform a joint task with seven other people one by one. After each joint task, they state whether they would like to work again with the person with whom they have just performed the task. Then, they play a final, payoff-relevant, round of the task with one of the people whom they stated they wanted to work with again. I find that the main determinant of partner selection is a given person's contribution to the task. However, after controlling for the contribution, people are significantly less likely to select a person who has corrected their actions. Women do not like being corrected both for their mistakes and for their right actions, while men mostly do not like being corrected for their mistakes. High ability men especially do not like to be corrected for their mistakes, suggesting that their emotional irritation is driving their negative reactions. The gender of the person who made corrections does not matter. I argue that these findings have implications for organizational efficiency, conflict, and gender differences in group work.

JEL codes: D91, C92, M54, J16
Keywords: correction, partner selection, group work, efficiency, gender differences

*PhD Candidate, Department of Economics, University of Bologna. Email: yuki.takahashi2@unibo.it. I am grateful to Maria Bigoni, Siri Isaksson, Bertil Tungodden, Laura Anderlucci, and Natalia Montinari whose feedback was essential for this project. I am also grateful to participants of the experiment for their participation and cooperation. Francesca Cassanelli, Alessandro Castagnetti, Mónica Costa-Dias, Ria Granzier-Nakajima, Annalisa Loviglio, Øivind Schøyen, Vincenzo Scrutinio, Erik Ø. Sørensen, Ludovica Spinola, and PhD students at the NHH and the University of Bologna all provided many helpful comments. This paper also benefited from participants' comments at the WEAI, PhD-EVS, and seminars at Ca' Foscari University, Catholic University of Brasília, the NHH, the University of Bologna, and the University of Copenhagen. Ceren Ay, Tommaso Batistoni, Philipp Chapkovski, Sebastian Fest, Christian König genannt Kersting, and oTree help & discussion group kindly answered my questions about oTree programming; in particular, my puzzle code was heavily based on Christian's code (https://github.com/chkgk/otree_slider_puzzle). 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

Being corrected by colleagues could improve group efficiency. For example, a research team published a paper in March 2020 arguing that a particular type of malaria treatment drug called hydroxychloroquine was effective for COVID-19. However, in May 2020, another researcher pointed out several flaws in the analysis and showed that the drug was not effective for COVID-19 (Bik 2020). If this scientist did not correct the flaws, some countries may have used the drug and COVID-19 patients may have suffered from the drugs' side effects.[1] However, being corrected by others could make people irritated rather than appreciating the correction. This could be very costly for a person who corrects others because she may reduce the probability of being selected into teamwork, which could be important for her career success.[2] For example, the hydroxychloroquine paper correction ended up the research team and its supporters attacking the researcher who pointed out the flaws (Davey 2021).

This paper studies how being corrected by others in a group affects one's probability of selecting that person as a partner in later works. To answer this question, I design a quasi-laboratory experiment, a hybrid of physical laboratory and online experiments. In the experiment, participants are grouped into people of eight, paired with another group member, solve one joint task together by alternating their moves. After solving the task, participants state whether they would like to be paired again with the same group member for the same task in the next stage, which is the main source of earnings. This gives a strong incentive for participants to select as good a partner as possible. Participants are paired with all the seven group members in a random order to address endogenous group formation. As a joint task, I use Isaksson (2018)'s number-sliding puzzle which allows me to calculate an objective measure of each participant's contribution to the joint task as well as to classify each move as good or bad.[3] I define a correction as reversing a group member's move.

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

These findings have implications to three strands of literature. The first strand of literature is social incentives in an organization (Ashraf and Bandiera 2018). Literature argues that managers favor workers whom they like in compensation and promotion (MacLeod 2003; Prendergast and Topel 1996) and workers tend to conform their managers (Prendergast 1993) when objective worker performance measures are not available, both of which distorts the optimal allocation of talent.[4] In addition, Li (2020) finds that this managers' favoritism not only distorts the optimal

1. See FDA (2020) for hydroxychloroquine's side effects.
2. Consider two assistant professors, one with several collaboration works with her colleagues and the other without collaboration works. When they face tenure evaluation, the former assistant professor is likely to have a better publication record than the latter assistant professor and more likely to get tenure.
3. Participants solve a 3x3 number-sliding puzzle in pairs by alternating their moves.
4. Several studies empirically verify these arguments in various different settings (Bandiera, Barankay, and

allocation of talent but also reduces non-favored workers' performance. My findings suggest that people's reluctance to accept being corrected can be a source of managers' favoritism and workers' conformity to managers. Also, the distortion may be larger when the manager is male because men are mostly reluctant to being corrected for their mistakes while women are reluctant to being corrected for their mistakes as well as their right actions.

The second strand of literature is psychology of conflict. Literature finds that people tend to view others who disagree with them as biased (Kennedy and Pronin 2008) and as having immoral motives (Reifen Tagar 2014). My findings that people react negatively to being corrected can be another source of conflict.

My findings also have implications to gender differences in corrections in group work. Guo and Recalde (2020) find that group members correct women's ideas more often than men's ideas. Isaksson (2018) finds that men are more likely to correct their group member's bad moves in the same puzzle used in my experiment.[5] I enrich this literature by showing that men are more prone to emotional irritations when being corrected by others.

2 Experiment

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

Introducing a quasi-laboratory format I run the experiment in a quasi-laboratory format where we experimenters connect us to the participants via Zoom throughout the experiment (but turn off participants' camera and microphone except at the beginning of the experiment) and conduct it as we usually do in a physical laboratory but participants participate remotely using their own computers.[6] On top of logistical convenience and complying with the COVID precaution measures, this quasi-laboratory format has an additional benefit over physical laboratory experiments that participants cannot see each other when they enter the laboratory which adds an additional layer of anonymity among participants. A drawback is that participants can be distracted while participating.

However, unlike standard online experiments such as on MTurk and Prolific where participants' identity is fully anonymous by the platforms' rule, we have participants' personal information and participants know it as we recruit them from our standard laboratory subject pool. Also, they are connected to us via Zoom throughout the experiment. These mostly prevent participants' attrition that can be endogenous to their decisions or treatments and the main problem of

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

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

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

online interactive experiments (Arechar, Gächter, and Molleman 2018) and experiments where treatments affect the probability of attrition, e.g., experiments with intertemporal decision making. In my experiment, we experienced no participant attrition. A drawback is that we could not collect a large number of observations.

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

Group task As the group task I use Isaksson (2018)'s puzzle, a sliding puzzle with 8 numbered tiles, which should be placed in numerical order within a 3x3 frame (see figure 3 for an example). To achieve this goal, participants play in pairs, alternating their moves. This puzzle has nice mathematical properties that I can define the puzzle difficulty and one's good and bad moves by the Breadth-First Search algorithm, from which I can calculate individual contributions to the group task and the quality of corrections objectively and comparably.[7] Further, the puzzle-solving captures an essential characteristic of group work in which two or more people work towards the same goal (Isaksson 2018) but the quality of each move and correction is only partially observable to participants (but fully observable to the experimenter).

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

Figure 1: Flowchart of the experiment

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

2.1 Design and procedure

Registration

Upon receiving an invitation email to the experiment, participants register for a session they want to participate in and upload their ID documents as well as a signed consent form.[8]

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

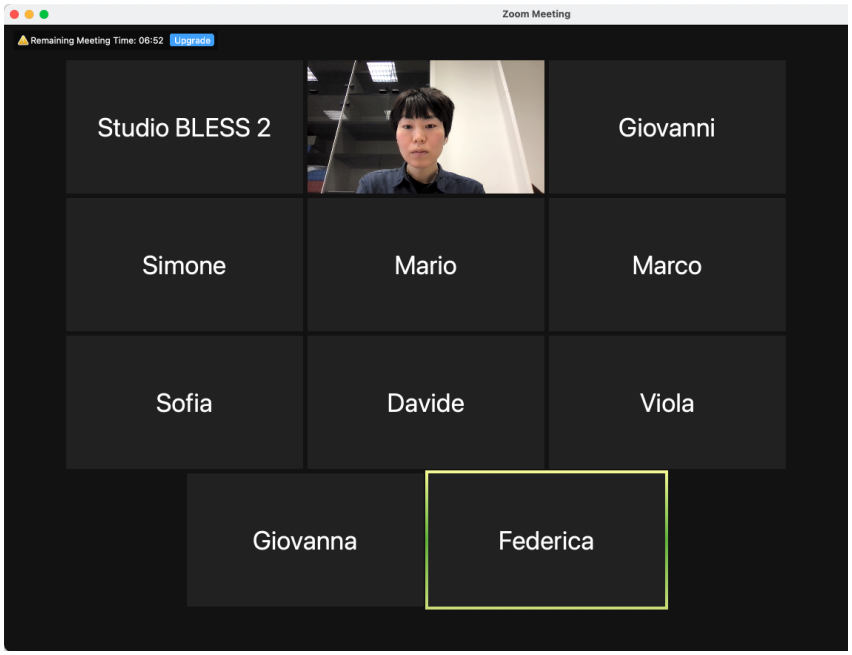
8. I recruit a few more participants than I would need for a given session in case some participants would not show up to the session.

Pre-experiment

On the day and the time of the session they have registered for, participants enter the Zoom waiting room.[9] 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 take attendance by calling each participant's first name one by one and ask her or him to respond via microphone. This process ensures other participants that the called participant's first name corresponds to her or his gender. If there are more participants than I would need for the session (I need 16 participants), I draw random numbers from 1 to 40 and ask those who drew the coins with the same number to leave.[10] Those who leave the session receive the 2€ show-up fee. Figure 2 shows a Zoom screen participants would see during the roll call (the person whose camera is on is the experimenter; participants would see this screen throughout the experiment but the experimenter's camera may be turned off).

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

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

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

Part 1: Solve puzzles individually

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

Part 2: Select a partner

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

Second, participants are paired with another randomly chosen participant in the same group and solve one puzzle together by alternating their moves. The participant who makes the first move is drawn at random and both participants know this first-mover selection criterion. If they cannot solve the puzzle within 2 minutes, they finish the puzzle without solving it. Participants are allowed to reverse the paired participant's move.[11] Each participant's performances in a given puzzle are measured as defined in Appendix A. Figure 3 shows a sample puzzle screen where a participant is paired with another participant called Giovanni and waiting for Giovanni to make his move.

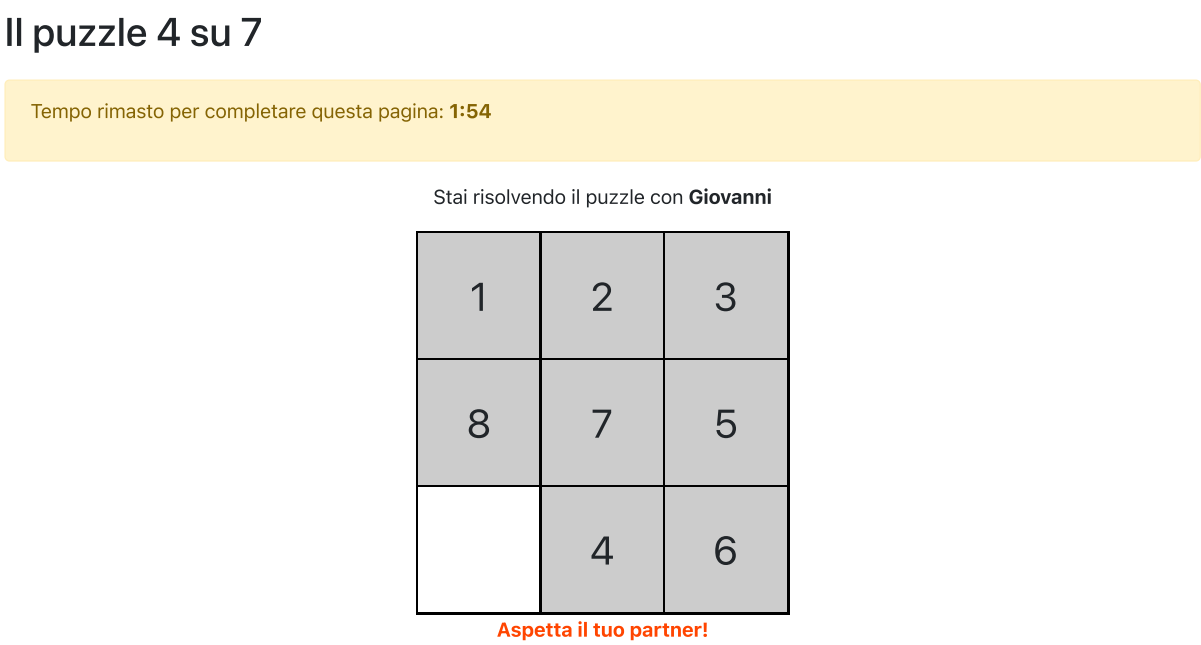
Once they finish the puzzle, participants state whether they would like to be paired again with the same participant in part 3 (yes/no). At the end of the first round, new pairs are formed, with a perfect stranger matching procedure, so that every participant is paired with each of the other 7 members of their group once and only once. In each round, participants solve another puzzle in a pair, then state whether they would like to be paired again with the same participant in part 3. The sequence of puzzles is the same for all pairs in all sessions. The puzzle difficulty is kept the same across the seven rounds. The minimum number of moves to solve the puzzles is set to 8 based on the pilot.

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

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

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

Figure 3: Puzzle screen



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

Figure 4: Partner selection screen

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

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

Part 3: Solve puzzles with a partner

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

Post-experiment

Each participant answers a short questionnaire which consists of (i) the six hostile and benevolent sexism questions used in Stoddard, Karpowitz, and Preece (2020) with US college students and (ii) their basic demographic information and what they have thought about the experiment. The answer to the sexism questions is used to construct a gender bias measure (see Appendix B for the construction of the measure) and their demographic information is used to know participants' characteristics as well as casually check whether they have anticipated that the experiment is about gender.[12]

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).[13] The first two conditions were to reduce noise coming from differences in socio-demographic backgrounds and race or/and ethnicity that may be inferred from participants' first name or/and voice and the last condition was to reduce experimenter demand effects. The number of participants was determined by a power simulation in the pre-analysis plan to achieve 80% power.[14] The experiment are pre-registered with the OSF.[15]

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

3 Data

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

12. None has anticipated that the puzzle is about gender.
13. The laboratory prohibits deception, so no participant has participated in an experiment with deception.
14. 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.
15. The pre-registration documents are available at the OSF registry: https://osf.io/tgyc5.

3.1 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 southern Italy are slightly overrepresented for both female and male participants.[16] 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).

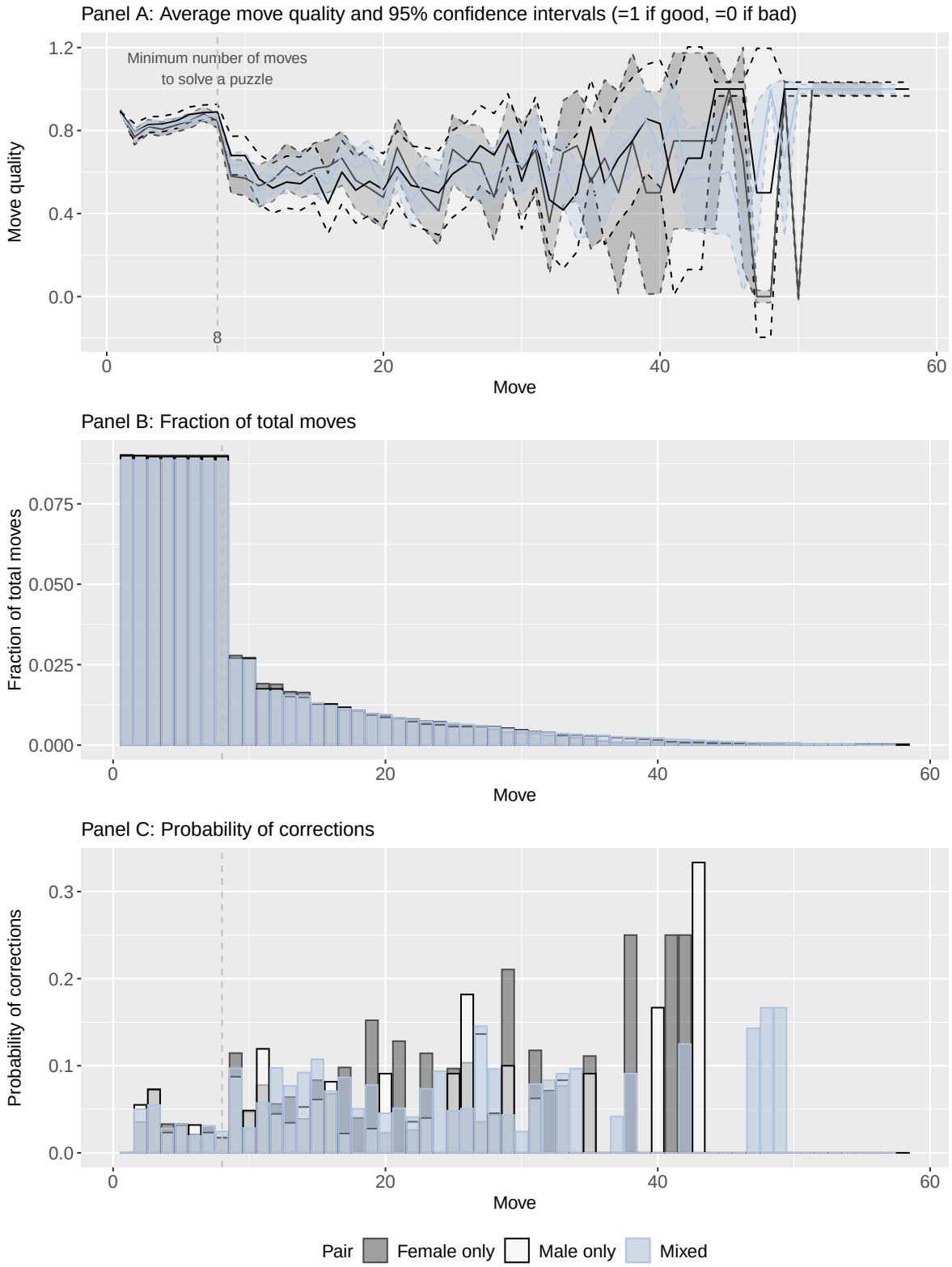
3.2 Move-level summary

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

Panel A shows that for all kinds of pairs, the average move quality is around 0.8 (8 out of 10 are good moves) until the 8th move (the minimum number of moves to solve a puzzle). After the 8th move, move quality deteriorates and stays around 0.6 (6 out of 10 are good moves).

16. Despite that I recruited only Italy-born people, 1 male participant answered in the post-questionnaire that he was from abroad. I include this participant in the analysis anyway but the results are robust to excluding this participant from the data.

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



Notes: The average move quality along with 95% confidence intervals (panel A), the fraction of total moves in each move (panel B), and the probability of corrections in each move (panel C), separately for female only (gray), male only (white), and mixed gender pairs (blue). The confidence interval of panel A is 95% confidence intervals of βs from the following OLS regression: $MoveQuality_{ijt} = \beta_1 + \sum_{k=2}^{58} \beta_k \mathbb{1}[t_{ij} = k] + \epsilon_{ijt}$, where t_{ij} is the pair i-j's move round and $\mathbb{1}$ is 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.

Panel B shows that for all kinds of pairs, about 71% of the puzzles are solved within 8 moves ((0.0875-0.025)/0.0875≈0.71), which is the minimum number of moves to solve the puzzle, then the other 30% takes more. Panel C shows that corrections happen across the moves, but are more likely to happen after the 8th move.

3.3 Puzzle-level summary

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

	Female (N=1708)		Male (N=1540)		Difference (Female – Male)		
	Mean	SD	Mean	SD	Mean	SE	P-value
Panel A: Own ability							
Contribution	0.44	0.19	0.45	0.18	-0.01	0.01	0.16
# puzzles solved in pt. 1	8.36	2.41	8.80	2.34	-0.44	0.22	0.05
Panel B: Partner's ability							
Contribution	0.45	0.19	0.45	0.18	0.00	0.01	0.55
# puzzles solved in pt. 1	8.58	2.35	8.57	2.43	0.01	0.16	0.93
Panel C: Corrections							
Being corrected	0.16	0.37	0.15	0.36	0.01	0.01	0.51
Being corrected a mistake	0.11	0.31	0.10	0.31	0.01	0.01	0.59
Being corrected a right move	0.05	0.21	0.05	0.21	0.00	0.01	0.77
Panel D: Puzzle outcomes							
Want to work again (yes=1, no=0)	0.72	0.45	0.71	0.45	0.01	0.02	0.49
Time spent (sec.)	43.74	36.15	42.99	35.76	0.74	1.28	0.56
Total moves	11.18	7.46	11.21	7.70	-0.03	0.28	0.92
Puzzle solved	0.85	0.36	0.86	0.35	-0.01	0.01	0.43
Consecutive correction	0.04	0.20	0.04	0.21	0.00	0.01	0.81

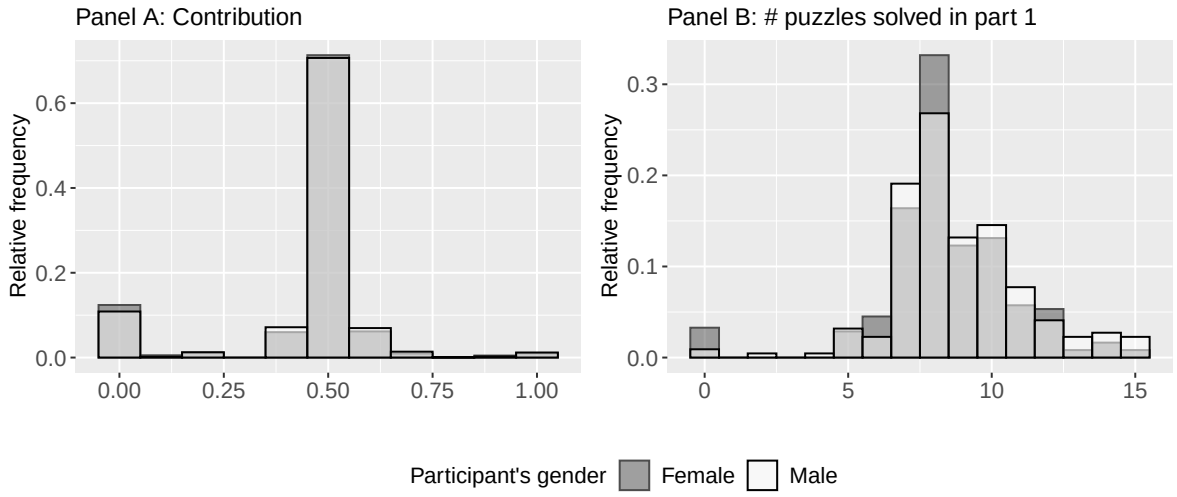
Notes: This table describes own (panel A) and partner's puzzle-solving ability (panel B), corrections received (panel C), and puzzle outcomes (panel D). P-values of the difference between female and male participants are calculated with CR0 standard errors clustered at the individual level. Appendix A provides definitions of each puzzle-solving ability measure.

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

Figure 6 shows distribution of ability measures to elaborate panel A of Table 2. First, panel A shows that both participants have contributed equally in about 70% of the puzzles. Second, male participants seem to have solved more puzzles even in distribution, which may be reflected in their puzzle moves in part 2 not captured by contribution, for example, speed of making a move.

Panel C shows that participants are corrected by their partner in 15-16% of the total puzzles,

Figure 6: Distribution of puzzle-solving ability



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

of which 10-11% are corrections of mistakes and 5% are corrections of a right move.[17]

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

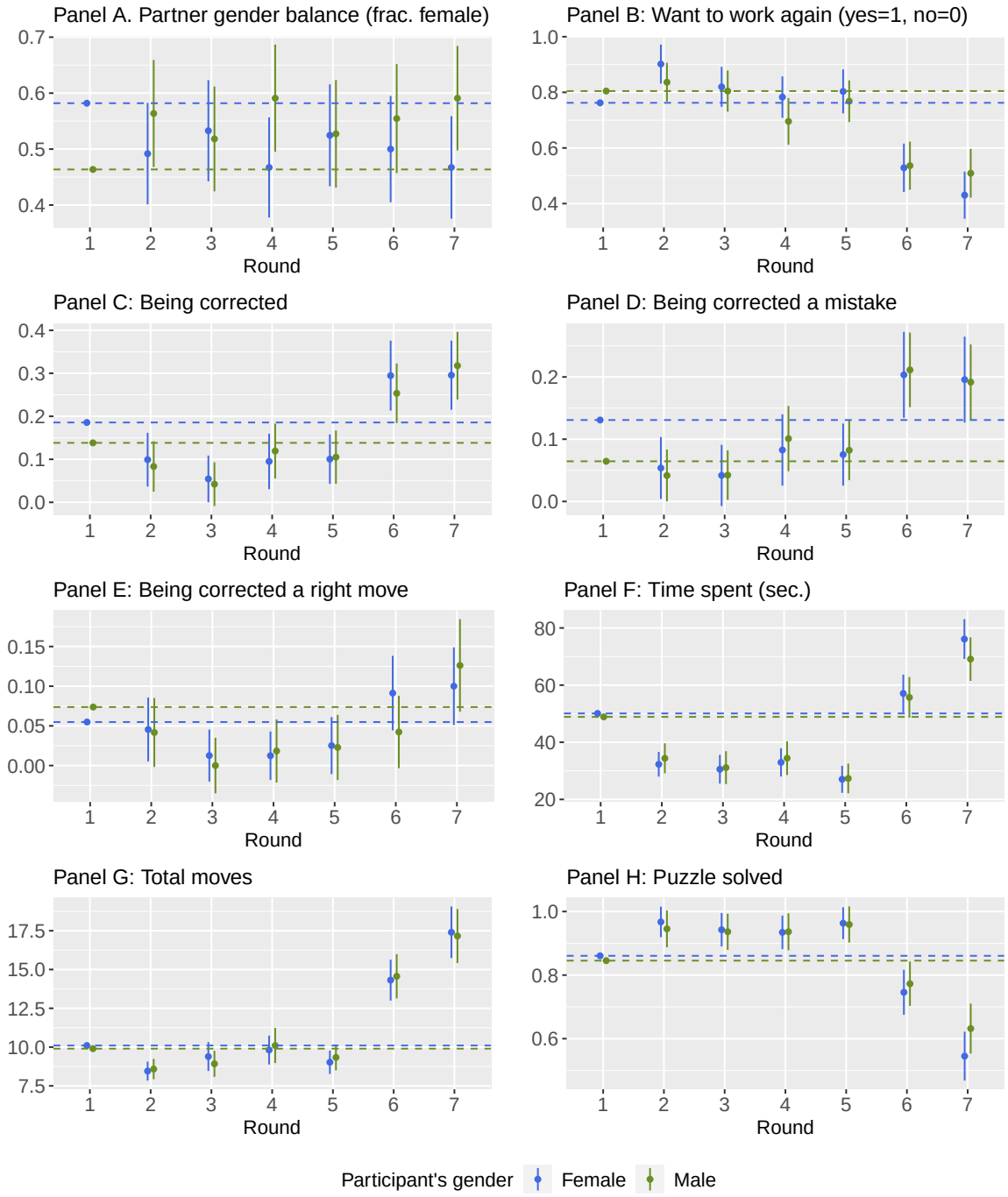
3.4 Balance across rounds

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

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

17. Of the 3180 puzzle, there are 495 puzzles where at least one correction occurred, of which 325 puzzles experienced only good corrections and 110 only bad corrections. The remaining 60 puzzles experienced both good and bad corrections. In order for good and bad corrections to capture only good and bad correction effect, I classify these 60 puzzles to good corrections if there were more good corrections than bad corrections (19 puzzles) and to bad corrections otherwise (41 puzzles).

Figure 7: Balance across rounds



Notes: This figure shows point estimates and 95% confidence intervals of βs from the following OLS regression with gender balance (female dummy) and different puzzle outcomes: $y_{ij} = \beta_1 + \sum_{k=2}^{7} \beta_k \mathbb{1}[t_{ij} = k] + \epsilon_{ij}$, where $t_{ij} \in \{1, 2, 3, 4, 5, 6, 7\}$ is the puzzle round in which i and j are playing, $\mathbb{1}$ is an indicator variable, and y_{ij} is outcome variable indicated in each panel. I add the estimate of β_1 to estimates of β_2-β_7 to make the figure easier to look at. The regression is estimated estimated separately for female (blue) and male participants (green). CR0 standard errors are clustered at the individual level.

rounds, puzzles in rounds 6 and 7 are perceived more difficult, etc.

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

4 Theoretical framework

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

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

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

where each term is defined as follows:

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

I assume:

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

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

4.1 When i is fully rational

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

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

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

4.2 When i is not fully rational

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

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

5 Response to being corrected

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

5.1 Response to being corrected: Estimating equation

I run the following OLS regression.

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

where each variable is defined as follows:

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

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

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

5.2 Response to being corrected: Results

Table 3 presents the regression results of equation 2. Columns 1, 3, and 5 show that when we do not control for partner's ability measures, correction effect is downward-biased, most likely

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

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

Table 3: Response to being corrected

Outcome:	Want to work again (yes=1, no=0)						
Sample:	All		Female		Male		All
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Being corrected	-0.367***	-0.269***	-0.406***	-0.314***	-0.322***	-0.214***	-0.313***
	(0.025)	(0.024)	(0.037)	(0.033)	(0.033)	(0.033)	(0.032)
Female partner	-0.006	0.012	-0.013	0.008	0.003	0.016	0.012
	(0.017)	(0.014)	(0.022)	(0.019)	(0.026)	(0.022)	(0.014)
Partner's contribution		1.175***		1.164***		1.192***	1.177***
		(0.054)		(0.075)		(0.076)	(0.054)
Partner's # puzzles solved in pt. 1		0.013***		0.018***		0.008	0.013***
		(0.004)		(0.006)		(0.006)	(0.004)
Being corrected x Male							0.096**
							(0.043)
Individual FE	✓	✓	✓	✓	✓	✓	✓
Baseline mean	0.780	0.780	0.780	0.780	0.778	0.778	0.780
Baseline SD	0.414	0.414	0.414	0.414	0.416	0.416	0.414
Adj. R-squared	0.076	0.312	0.078	0.325	0.076	0.300	0.313
Observations	3180	3180	1670	1670	1510	1510	3180
Clusters	464	464	244	244	220	220	464

Notes: This table presents regression results of equation 2 and shows that both women and men are less willing to work with a person who corrected their moves, but women respond stronger to being corrected. Columns 1, 3, and 5 excludes partner's puzzle-solving ability to show that coefficient estimate on being corrected is downward biased if we do not control for the ability. Baseline mean and standard deviation are that of partners who do not make corrections. CR0 standard errors in parentheses are clustered at the individual level. Significance levels: * 10%, ** 5%, and *** 1%.

because partners who make bad corrections tend to be low-ability and participants can partially observe move quality.

Looking at columns 2, 4, 6, and 7, the coefficient estimate on partner's contribution is positive and highly significant both quantitatively and statistically. This suggests that the main determinant of participant's partner selection is contribution to the puzzle. Coefficient estimate on partner's number of puzzles solved in part 1 is also statistically significant, but only for female participants.

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

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

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

6.1 A mistake vs. a right move: Estimating equation

I run the following OLS regression.

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

where each variable is defined as follows:

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

other variables are as defined in equation 2.

6.2 A mistake vs. a right move: Results

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

Outcome:	Want to work again (yes=1, no=0)						
Sample:	All		Female		Male		All
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Being corrected a mistake	-0.267***	-0.291***	-0.304***	-0.322***	-0.223***	-0.255***	-0.322***
	(0.031)	(0.028)	(0.046)	(0.037)	(0.040)	(0.042)	(0.038)
Being corrected a right move	-0.580***	-0.214***	-0.634***	-0.295***	-0.522***	-0.118**	-0.287***
	(0.036)	(0.036)	(0.048)	(0.050)	(0.052)	(0.048)	(0.046)
Female partner	-0.003	0.012	-0.011	0.008	0.006	0.014	0.011
	(0.017)	(0.014)	(0.022)	(0.019)	(0.026)	(0.022)	(0.014)
Partner's contribution		1.200***		1.172***		1.241***	1.202***
		(0.057)		(0.078)		(0.083)	(0.057)
Partner's # puzzles solved in pt. 1		0.014***		0.018***		0.008	0.013***
		(0.004)		(0.006)		(0.006)	(0.004)
Being corrected a mistake x Male							0.066
							(0.056)
Being corrected a right move x Male							0.157***
							(0.056)
Individual FE	✓	✓	✓	✓	✓	✓	✓
Being corrected a mistake	0.313***	-0.078*	0.330***	-0.026	0.299***	-0.138**	
=Being corrected a right move	(0.047)	(0.042)	(0.065)	(0.056)	(0.066)	(0.063)	
Baseline mean	0.780	0.780	0.780	0.780	0.778	0.778	0.780
Baseline SD	0.414	0.414	0.414	0.414	0.416	0.416	0.414
Adj. R-squared	0.092	0.312	0.096	0.324	0.089	0.302	0.314
Observations	3180	3180	1670	1670	1510	1510	3180
Clusters	464	464	244	244	220	220	464

Notes: This table presents regression results of equation 3 and shows that while women are less willing to work with a person who corrected their mistakes as well as right moves, men are mostly less willing to work with a person who corrected their mistakes. Baseline mean and standard deviation are that of partners who do not make corrections. CR0 standard errors in parentheses are clustered at the individual level. Significance levels: * 10%, ** 5%, and *** 1%.

Table 4 presents the regression results of equation 3. Looking at columns 1, 3, and 5 show that when we do not control for partner's ability measures, the effect of correction of a right move is severely downward biased. This is consistent with that participants can partially observe move quality.

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

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

7 Is negative response to being corrected rational or emotional?

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

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

7.1 Rational or emotional? Estimating equation

I run the following OLS regression.

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

where each variable is defined as follows:

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

other variables are as defined in equations 2 and 3.

7.2 Rational or emotional? Results

Table 5 presents the regression results of equation 4. Column 2 shows that high-ability women are neither less nor more reluctant to being corrected for their mistakes or for their right moves. However, column 4 shows that high-ability men are more reluctant to being corrected for their mistakes while their response to being corrected for their right moves are not different from low-ability men.

Table 5: A mechanism of negative response to being corrected

Outcome:	Want to work again (yes=1, no=0)			
Sample:	Female		Male	
	(1)	(2)	(3)	(4)
Being corrected a mistake	-0.289***	-0.292***	-0.188***	-0.192***
	(0.062)	(0.049)	(0.051)	(0.053)
Being corrected a right move	-0.636***	-0.279***	-0.526***	-0.099*
	(0.065)	(0.066)	(0.065)	(0.058)
Female partner	-0.011	0.008	0.006	0.014
	(0.022)	(0.019)	(0.026)	(0.022)
Partner's contribution		1.173***		1.249***
		(0.078)		(0.083)
Partner's # puzzles solved in pt. 1		0.018***		0.008
		(0.006)		(0.006)
Being corrected a mistake x High ability	-0.032	-0.063	-0.097	-0.180**
	(0.092)	(0.076)	(0.083)	(0.082)
Being corrected a right move x High ability	0.008	-0.034	0.015	-0.042
	(0.097)	(0.080)	(0.108)	(0.081)
Individual FE	✓	✓	✓	✓
Baseline mean	0.780	0.780	0.778	0.778
Baseline SD	0.414	0.414	0.416	0.416
Adj. R-squared	0.095	0.324	0.089	0.304
Observations	1670	1670	1510	1510
Clusters	244	244	220	220

Notes: This table presents regression results of equation 4 and shows that men's negative response to being corrected is due to their emotional irritation. Baseline mean and standard deviation are that of partners who do not make corrections. CR0 standard errors in parentheses are clustered at the individual level. Significance levels: * 10%, ** 5%, and *** 1%.

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

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

A study finds that men are motivated stereotyper that they view women in a (more) stereotypical way when women criticize them (Sinclair and Kunda 2000). Another study finds that people punish out-group members' misbehaviors more than that of in-group members (Chen and Li 2009). Thus, it may be that men respond more negatively to women's corrections than to men's corrections.

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

8.1 Does the gender of the partner matter? Estimating equation

I run the following OLS regression.

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

where each variable is defined as in equation 2.

8.2 Does the gender of the partner matter? Results

Table 6: Does the gender of the partner matter in response to being corrected?

Outcome:	Want to work again (yes=1, no=0)			
Sample:	Female		Male	
	(1)	(2)	(3)	(4)
Being corrected	-0.342***		-0.194***	
	(0.046)		(0.049)	
Being corrected a mistake		-0.367***		-0.214***
		(0.055)		(0.063)
Being corrected a right move		-0.280***		-0.144**
		(0.065)		(0.070)
Female partner	-0.001	-0.001	0.021	0.021
	(0.019)	(0.019)	(0.023)	(0.023)
Partner's contribution	1.164***	1.172***	1.189***	1.236***
	(0.075)	(0.077)	(0.077)	(0.083)
Partner's # puzzles solved in pt. 1	0.018***	0.018***	0.008	0.008
	(0.006)	(0.006)	(0.006)	(0.006)
Being corrected x Female partner	0.055		-0.039	
	(0.057)		(0.061)	
Being corrected a mistake x Female partner		0.091		-0.081
		(0.069)		(0.078)
Being corrected a right move x Female partner		-0.031		0.040
		(0.094)		(0.084)
Individual FE	✓	✓	✓	✓
Baseline mean	0.780	0.780	0.778	0.778
Baseline SD	0.414	0.414	0.416	0.416
Adj. R-squared	0.325	0.324	0.300	0.302
Observations	1670	1670	1510	1510
Clusters	244	244	220	220

Notes: This table presents regression results of equations 5 and shows that men do not respond more negatively to women's correction than to men's correction. Baseline mean and standard deviation are that of partners who do not make corrections. CR0 standard errors in parentheses are clustered at the individual level. Significance levels: * 10%, ** 5%, and *** 1%.

Table **??** presents the regression results of equation 5. Column 3 shows that men do not negatively respond to women's corrections than to men's correction. Column 4 shows that this is true even when we consider corrections of mistakes and of right moves.

While some of the coefficient estimates are negative, but women's coefficient estimates can

sometimes be negative (columns 1 and 2).

9 External validity and robustness

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

9.1 External validity

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

9.2 Robustness

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

To address this concern, I re-estimate equation 3 and 4 with solved puzzles only. Columns 1, 2, 5, and 6 of Table 7 present the results, which show that the findings are robust to excluding unsolved puzzles.

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

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

10 Discussion and conclusion

This paper studies how being corrected by others in a group affects one's probability of selecting that person as a partner in later works. I design a quasi-laboratory experiment where participants are paired with seven other participants, solve one number-sliding puzzle together, and express a preference on which of them to be paired with in the final, payoff-relevant, part of the experiment. I find that the paired participants' contribution to the puzzle is the most important factor for participants in selecting their partner. However, once I control for the paired participants' contribution to the puzzle, participants are significantly less likely to select a paired participant

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

Outcome:	Want to work again (yes=1, no=0)							
Sample:	Female, Solved puzzles		Female, Rounds 1-5		Male, Solved puzzles		Male, Rounds 1-5	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Being corrected a mistake	-0.328***	-0.302***	-0.278***	-0.223***	-0.390***	-0.301***	-0.323***	-0.211***
	(0.054)	(0.068)	(0.048)	(0.065)	(0.057)	(0.066)	(0.071)	(0.079)
Being corrected a right move	-0.219**	-0.138	-0.188**	-0.152	-0.018	-0.008	-0.103	-0.095
	(0.096)	(0.125)	(0.081)	(0.094)	(0.087)	(0.122)	(0.074)	(0.099)
Female partner	-0.001	-0.000	-0.001	-0.001	0.006	0.007	0.016	0.018
	(0.020)	(0.020)	(0.020)	(0.020)	(0.023)	(0.023)	(0.025)	(0.024)
Partner's contribution	1.411***	1.405***	1.400***	1.402***	1.724***	1.692***	1.337***	1.354***
	(0.259)	(0.259)	(0.130)	(0.129)	(0.263)	(0.260)	(0.132)	(0.128)
Partner's # puzzles solved in pt. 1	0.010*	0.010*	0.013**	0.013**	0.009	0.010*	0.013*	0.014**
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.007)	(0.007)
Being corrected a mistake x High ability		-0.046		-0.113		-0.203**		-0.358**
		(0.094)		(0.097)		(0.092)		(0.144)
Being corrected a right move x High ability		-0.166		-0.088		-0.026		-0.007
		(0.148)		(0.151)		(0.155)		(0.135)
Individual FE	✓	✓	✓	✓	✓	✓	✓	✓
Being corrected a mistake	-0.109		-0.090		-0.372***		-0.221**	
=Being corrected a right move	(0.125)		(0.095)		(0.119)		(0.105)	
Baseline mean	0.776	0.776	0.778	0.778	0.772	0.772	0.783	0.783
Baseline SD	0.417	0.417	0.416	0.416	0.420	0.420	0.412	0.412
Adj. R-squared	0.111	0.111	0.243	0.244	0.138	0.142	0.211	0.219
Observations	1449	1449	1199	1199	1321	1321	1083	1083
Clusters	244	244	244	244	220	220	220	220

Notes: This table presents regression results of equations 3 and 4 and shows that the findings so far are robust to exclusion of unsolved puzzles and rounds 6 and 7. Baseline mean and standard deviation are that of partners who do not make corrections. CR0 standard errors in parentheses are clustered at the individual level. Significance levels: * 10%, ** 5%, and *** 1%.

who corrected their move. Women do not like being corrected for their mistakes as well as their right moves, while men mostly do not like being corrected for their mistakes. High ability men especially do not like to be corrected for their mistakes, suggesting that the emotional irritation is driving their negative reactions. The gender of the paired participants who make corrections does not matter for negative response to being corrected.

These findings have three implications. First, people's reluctance to accept being corrected can be a source of managers' favoritism and workers' conformity to managers. Also, the distortion may be larger when the manager is male because men are mostly reluctant to being corrected for their mistakes while women are reluctant to being corrected for their mistakes as well as their right actions. Second, the finding that people react negatively to being corrected can be a source of conflict and speak to literature on psychology of conflict. Third, my findings on gender differences in response to being corrected enriches literature on gender differences in corrections in group work.

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.

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.

Bik, Elisabeth. 2020. *Thoughts on the Gautret et al. paper about Hydroxychloroquine and Azithromycin treatment of COVID-19 infections.* Science Integrity Digest.

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

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

Chen, Daniel L., Martin Schonger, and Chris Wickens. 2016. "oTree—An open-source platform for laboratory, online, and field experiments." *Journal of Behavioral and Experimental Finance* 9:88–97.

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

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

Davey, Melissa. 2021. "World expert in scientific misconduct faces legal action for challenging integrity of hydroxychloroquine study." *the Guardian.*

FDA. 2020. "FDA cautions against use of hydroxychloroquine or chloroquine for COVID-19 outside of the hospital setting or a clinical trial due to risk of heart rhythm problems." Drug Safety and Availability. https://www.fda.gov/drugs/drug-safety-and-availability/fda-cautions-against-use-hydroxychloroquine-or-chloroquine-covid-19-outside-hospital-setting-or.

Fisman, Raymond, Sheena S. Iyengar, Emir Kamenica, and Itamar Simonson. 2006. "Gender Differences in Mate Selection: Evidence from a Speed Dating Experiment." *The Quarterly Journal of Economics* 121 (2): 673–697.

———. 2008. "Racial Preferences in Dating." *The Review of Economic Studies* 75 (1): 117–132.

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

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

Greiner, Ben. 2015. "Subject pool recruitment procedures: organizing experiments with ORSEE." *Journal of the Economic Science Association* 1 (1): 114–125.

Guo, Joyce, and María P. Recalde. 2020. *Overriding in teams: The role of beliefs, social image, and gender.* Working Paper.

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

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

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

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

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

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

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

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

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

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

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.

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.

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

Appendix A Definition of performance measures

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

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

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

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

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

Appendix B Construction of the gender bias measure

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

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

> Instructions: Below is a series of statements concerning men and women and their relationships in contemporary society. Please indicate the degree to which you agree or disagree with each statement.
>
> 1. Women are too easily offended.
> 2. Many women are actually seeking special favors, such as hiring policies that favor them over men, under the guise of asking for "equality."
> 3. Men should be willing to sacrifice their own wellbeing in order to provide financially for the women in their lives.
> 4. Many women have a quality of purity that few men possess.
> 5. No matter how accomplished he is, a man is not truly complete as a person unless he has the love of a woman.
> 6. Women exaggerate problems they have at work.
>
> Answer choices to each question: Strongly agree, Agree a little, Neither agree nor disagree, Disagree a little, Strongly disagree

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

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.

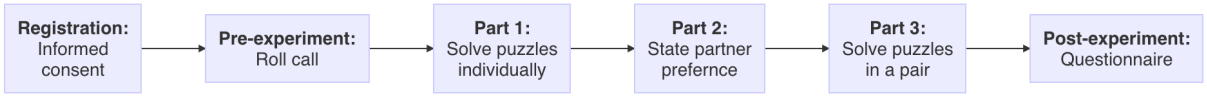
2 Design and procedure

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:

$$Prefer_{ij} = \beta_1 Contradict_{ij} * Female_j + \beta_2 Contradict_{ij} + \beta_3 Female_j + \delta Contribute_{ij} + IndividualFE_i + \epsilon_{ij} \quad (1)$$

- $Prefer_{ij} \in \{0, 1\}$: 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:

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

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:

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

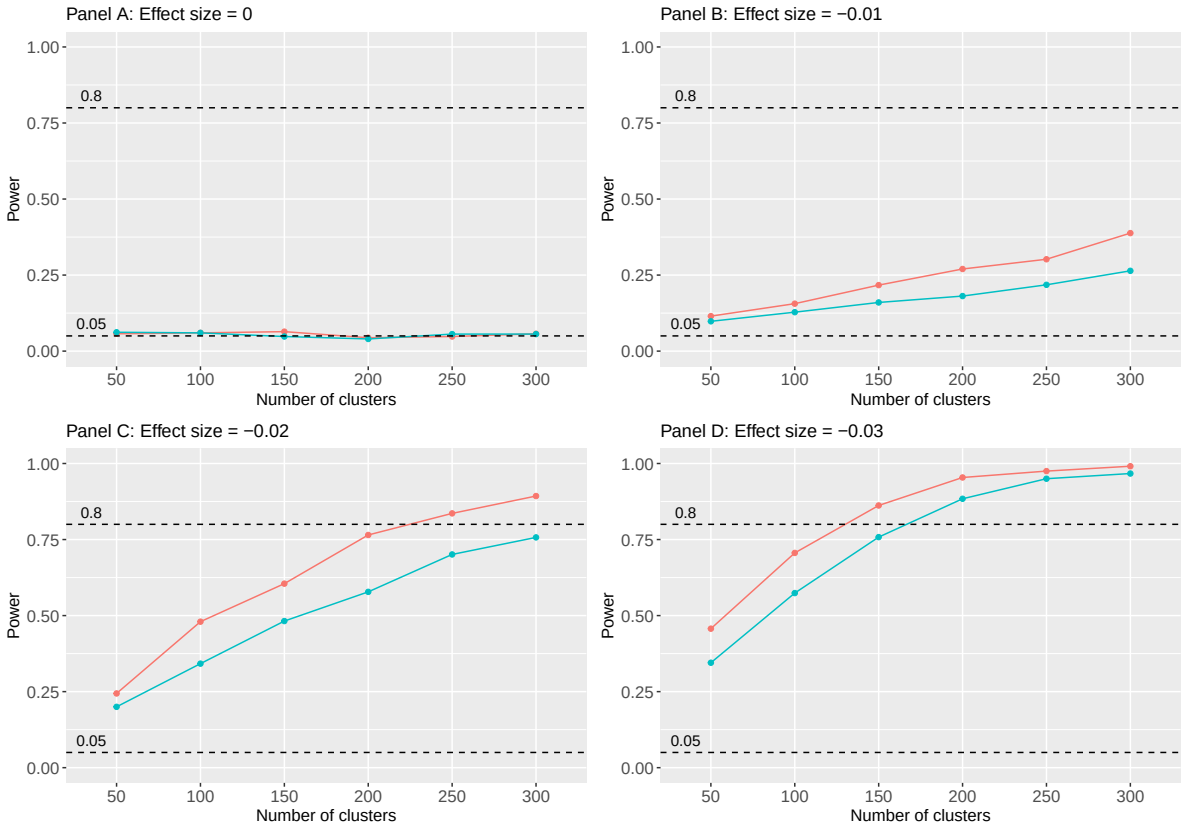
ATE is defined as the average of $\hat{\beta}_1$ across draws (its dependence on L is due to the non-linearity of the data generating process):

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

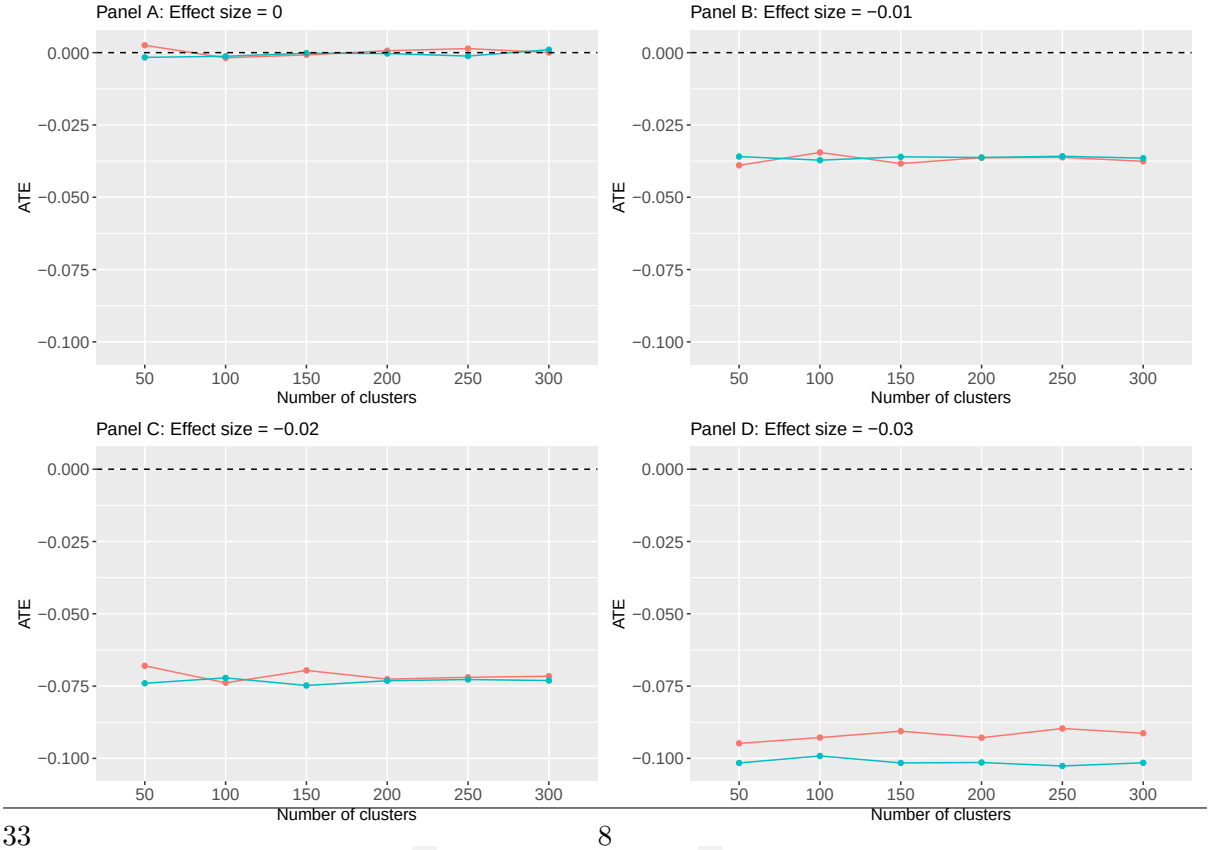
The results are presented in figure A1, which suggests that I need to recruit about 450 participants (so that I could have 225 clusters for testing H1). First, in the baseline scenario with $L = 15$, I can achieve about 80% power. Second, even under a tougher scenario where $L = 10$, I can still achieve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