Alma Mater Studiorum - Università di Bologna

DOTTORATO DI RICERCA IN

ECONOMICS

Ciclo 33

Settore Concorsuale: 13/A1 - ECONOMIA POLITICA

Settore Scientifico Disciplinare: SECS-P/01 - ECONOMIA POLITICA

ESSAYS IN GENDER ECONOMICS

Presentata da: Yuki Takahashi

Coordinatore Dottorato

Maria Bigoni

Supervisore

Maria Bigoni

Co-supervisori

Laura Anderlucci

Vincenzo Scrutinio

Esame finale anno 2022

Abstract

This PhD thesis is composed of three papers on gender economics.

The first chapter, **"Gender Differences in the Cost of Corrections in Group Work,"** studies whether people dislike collaborating with someone who corrects them and whether the dislike is stronger when that person is a woman. Having a good relationship with colleagues is integral in group work, potentially leading to successful collaborations. However, there are occasions when people have to correct their colleagues. Using a quasi-laboratory experiment, I find that people, including those with high productivity, are less willing to collaborate with a person who has corrected them even if the correction improves group performance. In addition, I find suggestive evidence that men respond more negatively to women's corrections, which is not driven by their beliefs about the difference in women's and men's abilities. These findings suggest that there is a behavioral bias in group work that distorts the optimal selection of talents and penalizes those who correct others' mistakes, and the distortion may be stronger when women correct men.

The second chapter, **"The Role of Gender and Cognitive Skills on Other People's Generosity,"** studies the role of gender and cognitive skills on other peoples' generosity. Cognitive skills are an important personal attribute that affects career success. However, colleagues' support is also vital as most works are done in groups, and the degree of their support is influenced by their generosity. Social norms enter in groups, and gender may interact with cognitive skills through gender norms in society. Because these gender norms penalize women with high potential, they can reduce colleagues' generosity towards these women. Using a novel experimental design where I exogenously vary gender and cognitive skills and sufficiently powered analysis, I find neither the two attributes nor their interactions affect other people's generosity; if anything, people are more generous to women with high potential. I argue that my findings have implications for the role of gender norms in labor markets.

The third chapter, **"The Welfare Effects of Increased Legal Tolerance toward Domestic Violence,"** studies how increased legal tolerance toward domestic violence affects married women's welfare using the domestic violence decriminalization bill introduced to the Russian national congress in 2016. Using difference-in-differences and flexibly controlling for macroeconomic shocks, I find that the bill decreased married women's life satisfaction and increased depression, especially among those with a college degree and a highly qualified white-collar occupation who are supposed to be more sensitive to gender regressive atmosphere. Consistent with this conjecture, people became more tolerant toward general and domestic violence after the bill. These findings suggest that the bill reduced married women's welfare partly through the gender regressive atmosphere.

Contents

Chapter 1

Gender Differences in the Cost of Corrections in Group Work

Yuki Takahashi*

Abstract

Having a good relationship with colleagues is integral in group work, potentially leading to successful collaborations. However, there are occasions when people have to correct their colleagues. I study whether people dislike collaborating with someone who corrects them and whether the dislike is stronger when that person is a woman. Using a quasi-laboratory experiment, I find that people, including those with high productivity, are less willing to collaborate with a person who has corrected them even if the correction improves group performance. In addition, I find suggestive evidence that men respond more negatively to women's corrections, which is not driven by their beliefs about the difference in women's and men's abilities. These findings suggest that there is a behavioral bias in group work that distorts the optimal selection of talents and penalizes those who correct others' mistakes, and the distortion may be stronger when women correct men.

JEL codes: J16, M54, D91, C92
Keywords: correction, collaboration, group work, gender differences, quasi-laboratory experiment

*Department of Economics, University of Bologna. Email: yuki.takahashi2@unibo.it. I am grateful to Maria Bigoni, Siri Isaksson, and Bertil Tungodden, whose feedback was essential for this project, and to the experiment participants for their participation and cooperation. This paper also benefited from helpful comments by Laura Anderlucci, Seda Ertaç, Valeria Ferraro, Boon Han Koh, Annalisa Loviglio, Valeria Maggian, Natalia Montinari, Vincenzo Scrutinio, participants at the CSQIEP Job Market Seminar, Stanford Institute for Theoretical Economics conference, Webinar in Gender and Family Economics, seminars at Ca' Foscari University, NHH, Osaka University, the University of Bologna, and many other people. Tommaso Batistoni, Philipp Chapkovski, 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. Francesca Cassanelli, Natalia Montinari, and Ludovica Spinola helped me to write experimental instructions in Italian. Michela Boldrini and Boon Han Koh conducted the quasi-laboratory experiments ahead of me and kindly answered my questions about the 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 on November 3, 2020 (ref. no. 262643).

1 Introduction

Having a good relationship with colleagues is vital in group work, and one key benefit is having successful collaboration opportunities. Since people do most work in groups both in the industry (Lazear and Shaw 2007) and in the academia (Jones 2021; Wuchty, Jones, and Uzzi 2007), successful collaboration opportunities are essential in their career.

However, on some occasions, one has to correct their colleague; one instance of such occasions is when a colleague at a business meeting suggests something against the group's interests (e.g., an investment plan that likely has a negative net present value) or interests of some group members (e.g., setting a meeting time later in the afternoon when nursery schools are already closed, which burdens people with small children). Another instance of such occasion is when a colleague presents their paper at an academic seminar which has some potential flaws in the identification strategy or the experimental design.

Such corrections may not affect one's relationship with their colleagues if those colleagues are rational agents,[1] but in reality, corrections may negatively affect one's relationship with their colleagues because people avoid information that conflicts with their own beliefs (Abelson 1986; Golman, Hagmann, and Loewenstein 2017). In particular, colleagues may dislike women's corrections more because people often use double standards for women and men (Egan, Matvos, and Seru 2021; Sarsons 2019; Sinclair and Kunda 2000).[2]

This paper studies whether people dislike collaborating with someone who corrects them and more so when that person is a woman. Answering this question using secondary data poses two challenges. First, group formation is not random and group corrections are endogenous. Second, different corrections are not necessarily comparable to each other.

To overcome these challenges, I design a quasi-laboratory experiment, a hybrid of physical laboratory and online experiments, where group formation is randomized and define corrections such that researchers can track its quality mathematically. Specifically, participants are allocated to a group of eight and solve one joint task with each group member one by one. Each time participants finish the task, they state whether they would like to collaborate with the group member with whom they have just solved the task for the same task in the next stage, which is the main source of earnings. This gives a strong incentive for participants to select as good a collaborator as possible. The order of the group members with whom participants solve the task is randomized. 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 (move the puzzle closer to the solution) or bad (move the puzzle further away from the solution). I define a correction as reversing a group member's move, which is comparable across different participants and can be classified as either good or bad.

1. Information avoidance is sometimes rational; see Golman, Hagmann, and Loewenstein (2017).

2. Evidence suggests that men undervalue women when they criticize them (Sinclair and Kunda 2000) and that people punish women more harshly when they make mistakes (Sarsons 2019) and commit misconduct (Egan, Matvos, and Seru 2021).

I find that people correctly understand the notion of good and bad moves; that is, the higher your contribution is to solving the puzzle, the more likely it is that you will be asked to join a team. This is in line with what one would expect and validates my experimental design.

Nonetheless, after controlling for the contribution, people are less willing to collaborate with a person who has corrected their moves, even if the corrections move the puzzle closer to the solution. Although it may not be so costly to correct colleagues if only low productivity people respond negatively to corrections because collaborations with those people will likely result in low quality outcomes, high productivity people also respond negatively to corrections. High productivity people's negative response also suggests that the negative response is likely to be irrational: they should be able to better identify good and bad corrections.

Although only suggestive, I also find evidence that men respond more negatively to women's good corrections: that is, men may dislike women correct their mistakes. This finding is unlikely to be due to people's beliefs about the differences in women's and men's abilities in the puzzle: women and men contribute equally well to the puzzle, and neither women nor men underestimate women's contribution.

Taken together, these findings suggest that there is a behavioral bias that distorts the optimal selection of talents and penalizes those who correct others' mistakes, and men may exhibit stronger bias when women correct them.

This paper's contribution is twofold. First, it contributes to the growing literature on the effect of workplace climate and productivity. For example, Alan, Corekcioglu, and Sutter (2021) find that a better workplace climate increases worker satisfaction and the degree of mutual reciprocation while reducing toxic competition and worker turnover, and argue that improved manager-worker relationships are the likely mechanism. Collaborating this evidence, Edmans (2011) finds that firms with high employee satisfaction exhibit higher stock prices, and Guiso, Sapienza, and Zingales (2015) find that a firm performs better when its workers perceive their managers as trustworthy and ethical. Also, workplace climate can differentially affect women and men: Dupas et al. (2021) find that female economists receive more patronizing and hostile questions during seminars. I show that interpersonal frictions can distort group efficiency in a controlled laboratory environment where I directly relate corrections and decisions to collaborate, and the friction may have a stronger effect on women.

Second, it contributes to the literature on gender differences in the contribution of ideas. For example, Coffman (2014) finds that women are less likely to contribute their ideas to the group in a male task due to self-stereotyping, and Gallus and Heikensten (2019) find that debiasing their self-stereotyping by giving an award for their high ability increases women's contribution: they put women's idea further ahead of men without involving open correction of their group member. However, on some occasions, one has to contribute their ideas openly, for example, at business meetings and academic seminars. In such cases, group members' response plays an important role in the effectiveness of the intervention suggested by Gallus and Heikensten (2019), and indeed there is evidence that people respond to women's ideas less favorably: Coffman, Flikkema, and Shurchkov

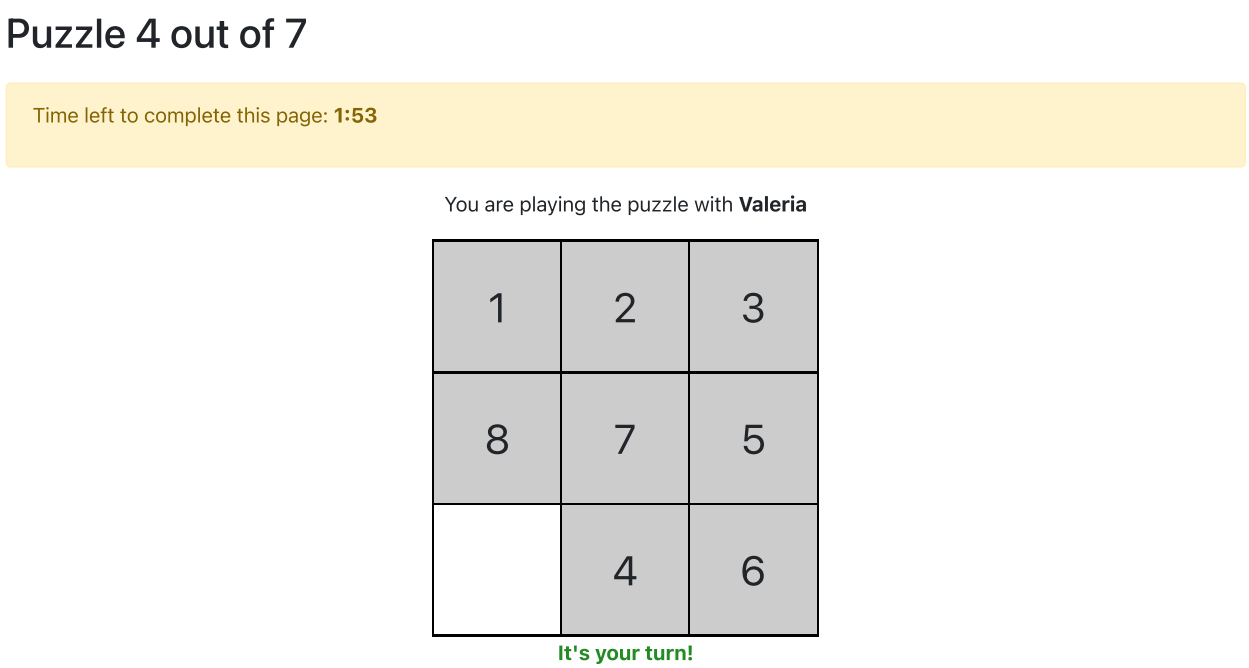
(2021) find that group members are less likely to choose women's answers as a group answer in male-typed questions. Corroborating this, Guo and Recalde (2022) find that group members correct women's ideas more often than men's ideas. I introduce correction in the contribution of ideas and examine its cost on women and groups so that we can design more effective interventions.

The remainder of the paper proceeds as follows. Section 2 describes the experimental design, procedure, and implementation. Section 3 describes the data. Section 4 presents a simple theoretical framework. Section 5 presents evidence that people are less willing to collaborate with a person who has corrected their moves, even if the corrections move the puzzle closer to the solution. Section 6 presents evidence that even high productivity people respond negatively to corrections. Section 7 presents suggestive evidence that men respond more negatively to women's good corrections. Section 8 presents the robustness of the results. Section 9 concludes.

2 Experiment

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

Figure 1: Puzzle screen



Notes: This shows a sample puzzle screen where a participant is matched with another participant called Valeria at the 4th round of the puzzle and making their move. All the texts are in Italian in the experiment.

Group task As the group task, I use Isaksson (2018)'s puzzle, a sliding puzzle with eight numbered tiles, which should be placed in numerical order within a 3x3 frame (see Figure 1 for an example). To achieve this goal, participants play in pairs, alternating their moves.[3] This puzzle has nice mathematical properties: I can define the puzzle difficulty and classify a given move as either good or bad by the Breadth-First Search algorithm.[4] From the number of good and bad moves one makes, I can calculate individual contributions to the group task; I measure it by net good moves, the number of good moves minus the number of bad moves an individual makes in a given puzzle.

I can also determine the quality of corrections of different participants objectively and comparably.[5] Further, the puzzle-solving captures an essential characteristic of teamwork 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).

At each stage of the puzzle, there is only one good strategy which is to make a good move and one bad strategy which is to make a bad move.[6] There can be more than one good and bad move, but different good/bad moves are equal. There is no path dependence either: the history of the puzzle moves does not matter.

At the beginning of each part, participants must answer a set of comprehension questions to make sure they understand the instructions.[7]

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]

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.

3. Each participant has to make a move in their turn; they cannot pass.

4. 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.

5. Indeed, some corrections happen early in the puzzle and the other later in the puzzle. Thus, what I capture in the analysis is the average effect of a correction.

6. This is conditional on that both players are trying to solve the puzzle; I show in section 8 that the results are robust to exclusion of puzzles where either player might not be trying to solve the puzzle.

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

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.

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.

Then I admit participants to the Zoom meeting room one by one and rename them by the first name they have just entered. This information is necessary to match up their earnings in this experiment and their payment information stored in the laboratory database, so participants have a strong incentive to provide their true name and email address. If there is more than one participant with the same first name, I add a number after their first name (e.g., Giovanni2).

After admitting all the participants to the Zoom meeting room, I do roll call, a way to reveal participants' gender to other participants without making gender salient (Bordalo et al. 2019; Coffman, Flikkema, and Shurchkov 2021). Specifically, 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 their 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).

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: Individual practice stage

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.

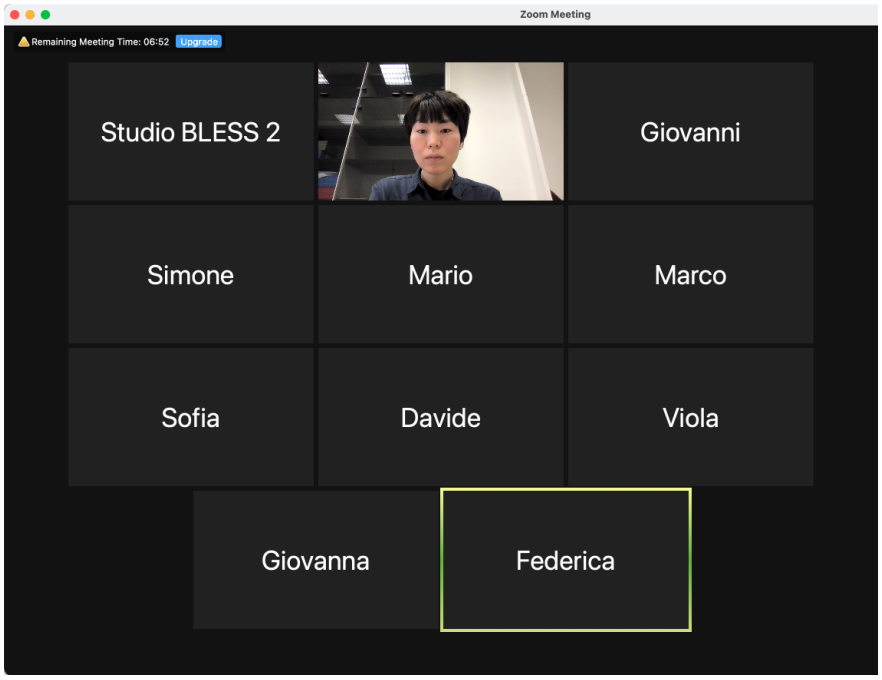
Part 2: Collaborator selection stage

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

10. I draw with replacement a number from 1 to 40 using Google's random number generator (https://www.google.com/search?q=random+number). 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.

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.

solve the puzzle within 2 minutes, they finish the puzzle without solving it. Participants are allowed to reverse the paired participant's move.[11] Reversing the partner's move is what I call correction in this paper. Each participant's contribution in a given puzzle is measured by net good moves. Figure 1 shows a sample puzzle screen where a participant is paired with another participant called Valeria and making their move.[12] The paired participant's first name is displayed on the computer screen throughout the puzzle and when participants select their collaborator to subtly inform the paired participant's gender.

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

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

12. All the texts are in Italian in the experiment.

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

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

Part 3: Group work stage

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

Post-experiment

Each participant answers a short questionnaire which consists of (i) the six hostile and benevolent sexism questions used in Stoddard, Karpowitz, and Preece (2020) with US college students and (ii) their basic demographic information and what they have thought about the experiment.[13] The answer to their demographic information is used to know participants' characteristics as well as casually check whether they have anticipated that the experiment is about gender, for which I do not find any evidence.

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 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).[14] The first two conditions were to reduce noise coming from

13. I was planning to use a gender bias measure constructed from the hostile and benevolent sexism questions to show those with higher gender bias respond more negatively to women's corrections. However, people do not respond more negatively to women's corrections and that I could not have enough variation in this gender bias measure, so decided not to report it in the main text; the results are reported in Appendix B.

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

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.[15] The number of participants was determined by a power simulation in the pre-analysis plan to achieve 80% power.[16] The experiment is pre-registered with the OSF.[17]

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. Table 1 describes participants' characteristics. The table shows that while female participants are slightly younger (1.41 years) and less gender biased (0.12). In addition, 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).[18] Also, most female and male participants are either bachelor or master students (97% of female and 94% of male) and only few are PhD students.

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 (within Italy)								
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
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. P-values of the difference between female and male participants are calculated with heteroskedasticity-robust standard errors.

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

16. 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.

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

18. Individual fixed effects in the analysis control for one's major. However, I do not run heterogeneity analysis by major because major choice is endogenous to one's gender.

3 Data description

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

Move-level data Figure 3 shows average move quality across moves along with 95% confidence bands (Panel A), fraction of total moves (Panel B), and probability that a correction is happening (Panel C), for female only pairs (blue), male only pairs (green), and mixed gender pairs (black-white). Panel A shows that there is no statistically significant differences in move quality by own gender or the gender of the partner. Panel B shows that about 71% of the puzzles are solved within a minimum number of moves (the minimum number of moves is 8) and shows that own gender or the gender of the partner does not matter in how fast participants solve the puzzle. Panel C shows that corrections happen across the moves, but there is no systematic differences in the probability that correction is happening by own gender or the gender of the partner.

Puzzle-level data Table 2 describes own (panel A) and partner's puzzle behaviors (panel B) and puzzle outcomes (panel C). Panel A shows that there are no gender differences in puzzle-solving ability: both contribution in part 2 and the number of puzzles solved in part 1, the difference between female and male participants are statistically insignificant at 5% and quantitatively insignificant.[19][20] This is consistent with Isaksson (2018), who also finds no gender difference in contribution or number of puzzles solved alone using the same puzzle, suggesting that any gender difference I would find is unlikely to come from their ability difference. Panel A also shows that there are no gender differences in propensity to correct partners, unlike Isaksson (2018), who finds that men correct their partner more often than women, although their result is from move-level data. Finally, the last row of Panel A shows that male participants are slightly more likely to face female partners, although only 3 percentage points more.

To further elaborate panel A of Table 2, Panel A of Figure 4 presents the distribution of contribution by participants gender that women and men are equally good at puzzle solving: in about 70% of the puzzles, participants' contribution is 4 (total good moves minus total bad moves), and women's and men's distributions almost overlap.

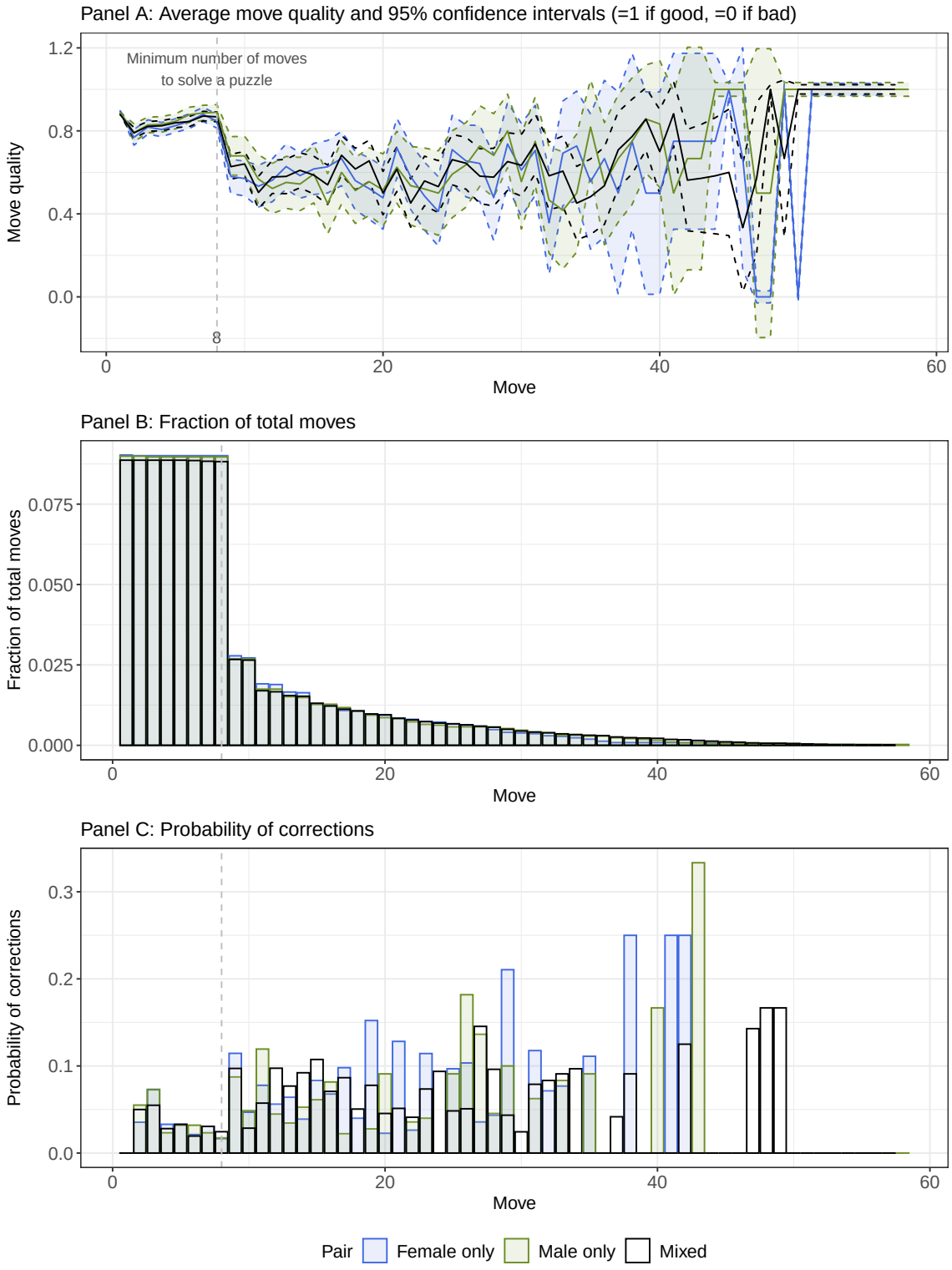
Panel B shows that puzzle-solving ability as well as propensity to make corrections (both of a mistake and of a right move) of partners paired with female and male participants is the same, suggesting random pairing was successful and that any gender differences I would find are not coming from partners of either gender correct more often. Participants are corrected by their partner in 15-16% of the total puzzles, of which 12-13% are good corrections, and 5-6% are bad corrections, and there are no gender differences in propensity to be corrected.[21]

19. The number of puzzles solved in part 1 is marginally significant but quantitatively insignificant.

20. The correlation coefficient between contribution and number of puzzles solved in part 1 is 0.1059 and the p-value is below 0.001 (with standard errors clustered at individual level).

21. The percentage of good corrections and bad corrections do not sum up to the percentage of any correction

Figure 3: Move quality, fraction of total moves, and probability of corrections



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

Table 2: Own and partners' puzzle behaviors and puzzle outcomes

	Female (N=1708)		Male (N=1540)		Difference (Female – Male)		
	Mean	SD	Mean	SD	Mean	SE	P-value
Panel A: Own behaviors							
Contribution	2.98	2.93	3.14	2.64	-0.16	0.10	0.11
# puzzles solved in part 1	8.36	2.41	8.80	2.34	-0.44	0.22	0.05
Any correction	0.15	0.36	0.16	0.36	0.00	0.01	0.85
Good correction	0.12	0.33	0.12	0.33	0.00	0.01	0.90
Bad correction	0.06	0.23	0.05	0.22	0.00	0.01	0.70
(Fraction of female partners)	0.51	0.50	0.54	0.50	-0.03	0.02	0.03
Panel B: Partner's behaviors							
Contribution	3.04	2.73	3.07	2.87	-0.03	0.10	0.77
# puzzles solved in part 1	8.58	2.35	8.57	2.43	0.01	0.16	0.93
Any correction	0.16	0.37	0.15	0.36	0.01	0.01	0.51
Good correction	0.13	0.33	0.12	0.32	0.01	0.01	0.44
Bad correction	0.06	0.23	0.05	0.22	0.01	0.01	0.44
Panel C: Puzzle outcomes							
Willing to collaborate (yes=1, no=0)	0.72	0.45	0.71	0.45	0.01	0.02	0.49
Time spent (second)	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

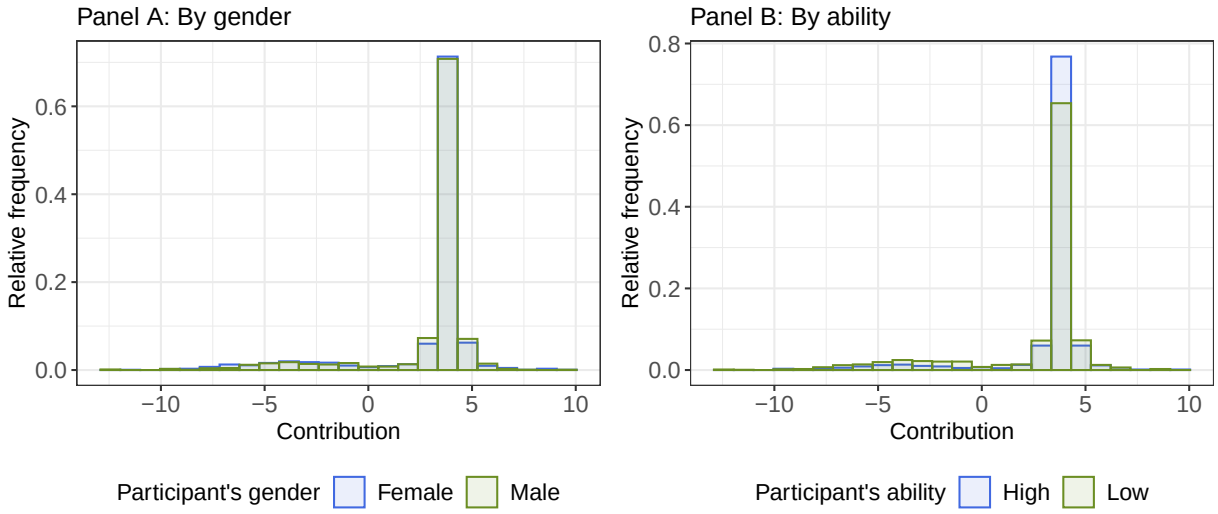
Notes: This table describes own (panel A) and partner's puzzle behaviors (panel B) and puzzle outcomes (panel C). P-values of the difference between female and male participants are calculated with standard errors clustered at the individual level. Contribution is defined as one's net good moves in a given puzzle (the number of good moves minus the number of bad moves).

To look more closely into whether there are really no gender differences in propensity to correct and to be corrected by partners that we saw in Panels A and B of Table 2, Figure 5 plots propensity to correct a female partner (Panel A), a male partner (Panel B), propensity to be corrected by a female partner (Panel C), and by a male partner (Panel D), separately for female and male participants and for any correction, good correction, and bad correction along with the 95% confidence intervals. The figure shows that female participants may be slightly more likely to make a bad correction to a male partner (Panel B), receive a good correction from a female partner (Panel C), and receive a bad correction from a male partner (Panel D), but none of them are statistically significant.

Panel C shows that participants state they want to collaborate with the partner 71-72% of the time. Participants spend on average 43-44 seconds for each puzzle (the maximum time a pair can spend is 120 seconds), and take 11 moves. 85-86% of the puzzles are solved and participants and the

means there are puzzles where both good and bad corrections occurred. The results are robust to exclusion of these overlapping puzzles, as shown in Figures 7, 8, and 9.

Figure 4: Distribution of contribution



Notes: This figure shows the distribution of individual contribution by gender (panel A) and ability (panel B) and shows that most participants contributed to the same degree. Panel A further shows no gender difference in contribution, and panel B further shows that among high-ability people, higher fraction contributes to the puzzles to the same degree. Contribution is defined as one's net good moves in a given puzzle (the number of good moves minus the number of bad moves).

partner correct each other's move consecutively in 4% of the puzzles.[22] There is no gender difference in any of these outcomes, suggesting any gender differences cannot be attributed to the imbalance in these outcomes.[23]

Across-round balance Figure 6 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 (relative to round 1), separately for female (blue) and male participants (green).

First, there is some unbalance in partner's gender across rounds between female and male participants (Panel A), with female/male participants more/less likely to be paired with female partner in round 1, but the difference is not statistically significant for rounds 2-7.

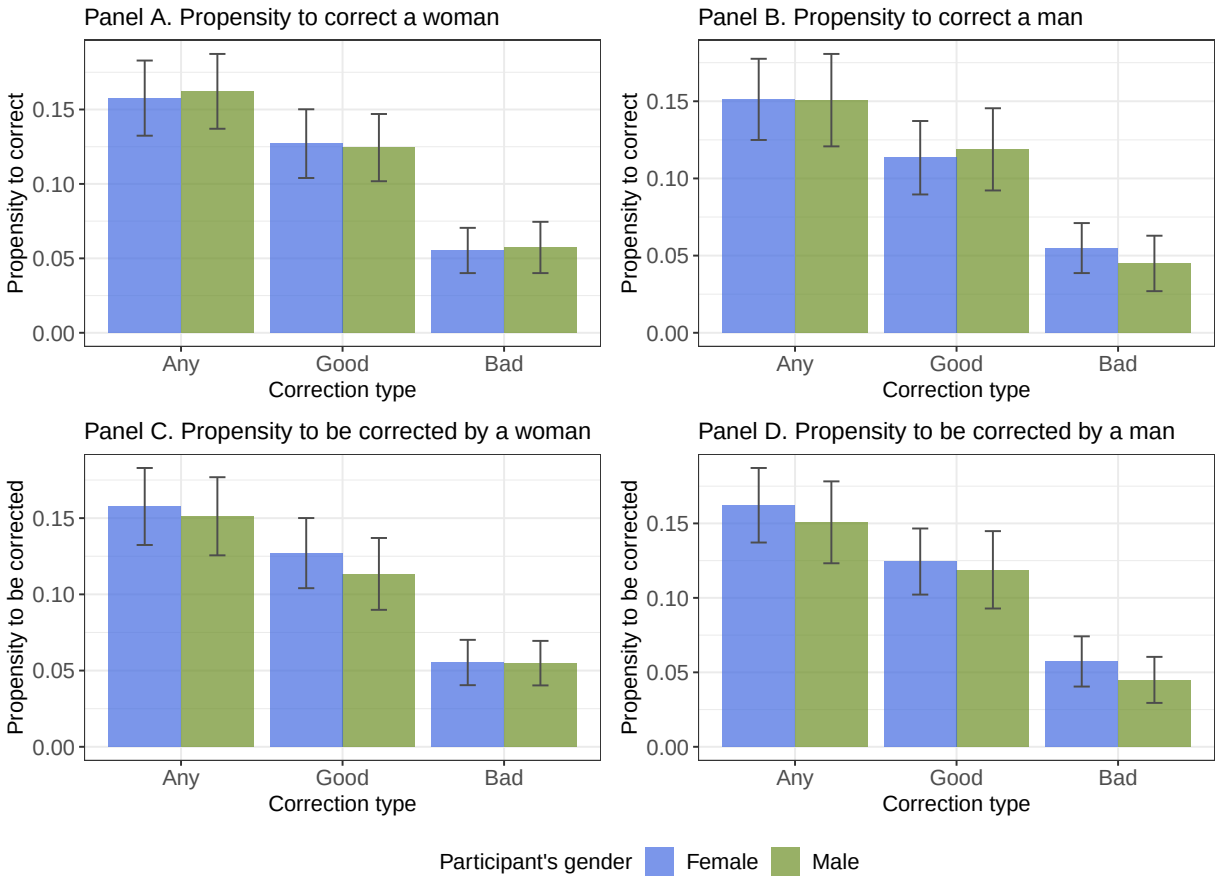
Second, we see that there is no systematic gender differences in puzzle outcomes across rounds (Panels B-H), suggesting that female and male participants behave similarly across rounds. One difference could be good and bad corrections, with female participants make slightly more bad corrections and slightly less good corrections. However, as shown in Table 2, these differences are statistically insignificant.

Last, we see that in rounds 6 and 7, participants are less willing to collaborate, experience more corrections, and less likely to solve the puzzle. Although they are all outcomes of a particular pair

22. Indeed, in puzzles where consecutive correction happens, probability of selecting a paired participant as collaborator drops from 78.0% to 26.8%.

23. Note that time spent to solve a puzzle is endogenous to correction and not a good control. For example. if one corrects a mistake, then it takes fewer time to solve the puzzle. If one corrects a right move, on the other hand, then it takes more time to solve the puzzle.

Figure 5: Propensity to correct/to be corrected



Notes: This figure plots propensity to correct a female partner (Panel A), a male partner (Panel B), propensity to be corrected by a female partner (Panel C), and by a male partner (Panel D), separately for female and male participants and for any correction, good correction, and bad correction along with the 95% confidence intervals.

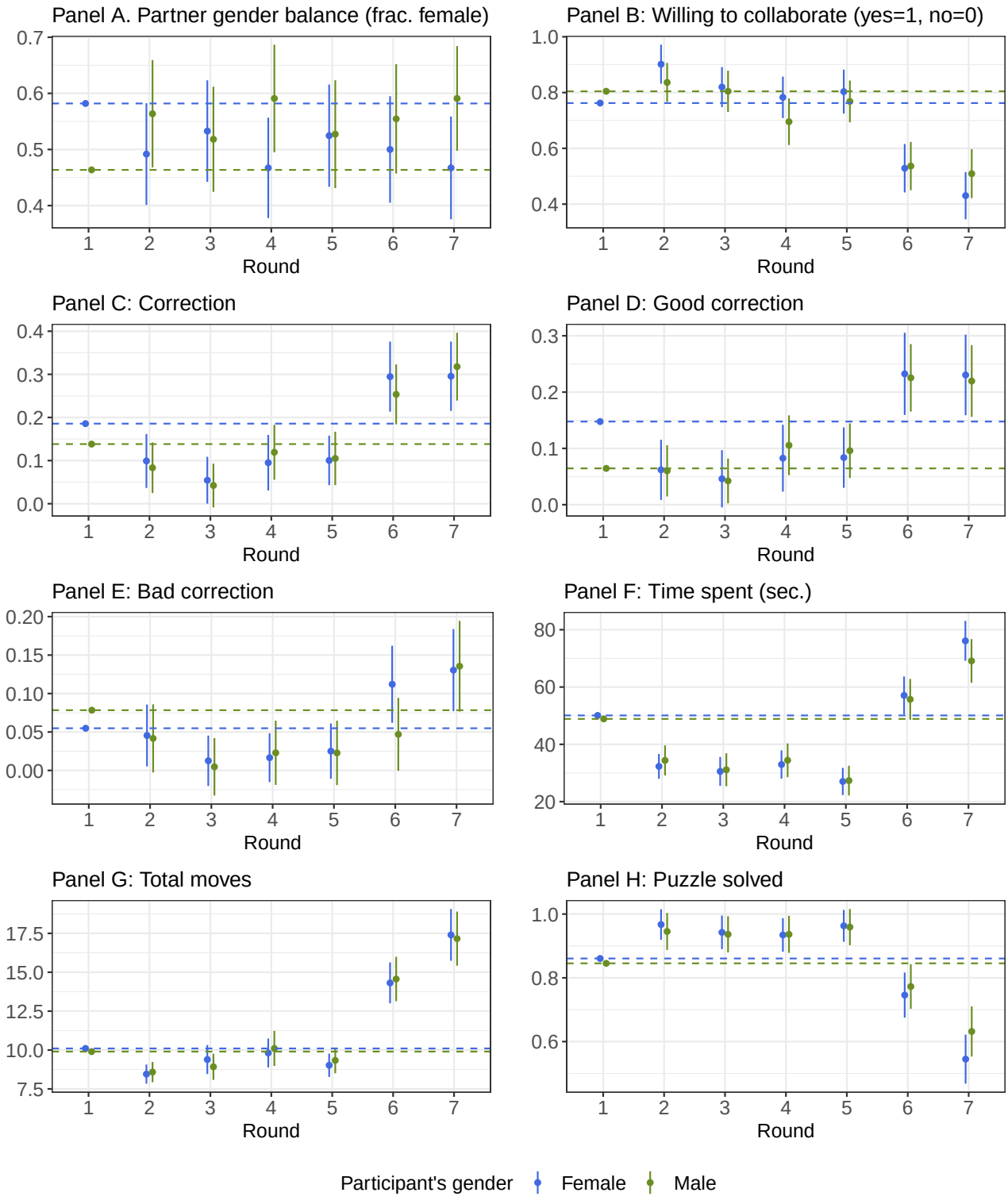
which is randomly formed, it can simply be correlations. Still, one may wonder whether rounds 6 and 7 are driving the results. I will show in section 8 that the results are robust to exclusion of these rounds.

4 Theoretical framework

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

I consider a fully rational participant i who maximizes their expected payoff in a given round t by deciding whether they are willing to collaborate with a potential collaborator j with whom they have just played one puzzle, conditional on the history of decisions i has made to other potential collaborators with whom they have played the puzzle up to the current round t and with whom they will play the puzzle in the future rounds. Since with whom to be paired in which order is randomized, I simply denote the history and the future by t, consider them as exogenous, and normalize the payoff of not willing to collaborate with j as 0 for each round t.

Figure 6: Balance across rounds



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

The payoff is increasing in i's belief about j's ability. I assume i can partially observe j's move quality, so i's belief about j's ability is increasing in j's ability perceived by i.

Thus, i would face the following problem:

$$\max_{Accept \in \{0,1\}} \mathbb{1}[Accept = 1] \times E_{\mu_j}[\pi_t(\mu_j(\tilde{a}_j, c_j, f_j))|\theta, \omega, t], \quad \partial \pi_t / \partial \mu_j > 0, \ \partial \mu_j / \partial \tilde{a}_j > 0 \tag{1}$$

where each term is defined as follows:

- *Accept*: whether i is willing to collaborate with j (=0 if no, =1 if yes)
- 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 if j did not correct i)
- f_j: j's gender (=1 if female, =0 if male)
- θ: i's belief about their ability relative to other participants in the session (>0 if higher, =0 if same, <0 if lower)
- ω: j's belief about women's ability relative to men (>0 if higher, =0 if same, <0 if lower)

where $\mathbb{1}$ is an indicator function. Although θ and ω could depend on t, I omit the dependence on t for simplicity because t is exogenous.

If i can fully observe j's move quality and i is fully rational, then j's correction, c_j, and gender, f_j, do 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 and gender convey information about j's ability even if i is fully rational.[24]

First, keeping j's ability perceived by i fixed, the information j's correction conveys depends on θ. If i believes they are good at the puzzle, they would consider a correction as a signal of low ability because i believes their move is correct. On the other hand, if i believes their ability is low, then they would consider a correction as a signal of high ability. If i believes their ability is the same as j's, then a correction would not convey any information.

However, since i can partially observe j's move quality, i considers a good correction as less negative/more positive signal than a bad corrections regardless of θ. Thus, we have the following proposition:

Proposition 1. *A fully rational participant i is less willing to collaborate with j when j made a bad correction than when j made a good correction, regardless of i's belief about their own ability. That is:*

$$\partial \mu_j / \partial c_j|_{c_j \text{ is a bad correction}} < \partial \mu_j / \partial c_j|_{c_j \text{ is a good correction}} \ \forall \theta \tag{2}$$

Also, the more the i understands the puzzle, the more they can observe j's move quality, hence corrections, regardless of θ. Thus, we have the following proposition:

Proposition 2. *A fully rational participant i with higher puzzle solving ability is more willing to collaborate with j when j made a good correction and less willing to collaborate with j when j made*

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

a bad correction, compared to another fully rational participant with lower puzzle solving ability. This is true regardless of their belief about their ability. That is:

$$\begin{aligned} \partial\mu_j/\partial c_j|_{i\text{'s ability is higher} \wedge c_j \text{ is a good correction}} &> \partial\mu_j/\partial c_j|_{i\text{'s ability is lower} \wedge c_j \text{ is a good correction}} \ \forall\theta \\ \partial\mu_j/\partial c_j|_{i\text{'s ability is higher} \wedge c_j \text{ is a bad correction}} &< \partial\mu_j/\partial c_j|_{i\text{'s ability is lower} \wedge c_j \text{ is a bad correction}} \ \forall\theta \end{aligned} \tag{3}$$

Similar to response to corrections, if i believes women is better at the puzzle, they would consider a correction from a woman as a signal of high ability relative to men's correction. On the other hand, if i believes women is worse at the puzzle, then they would consider a correction from a woman as a signal of low ability relative to men's correction. If i believes women and men are equally good at the puzzle, then a correction from a woman or man is irrelevant. Thus, we have the following proposition:

Proposition 3. *A fully rational participant i is willingness to collaborate with j when j was a woman and made a correction relative to when j was a man and made a correction depends on their belief about women's ability relative to men's. This is true regardless of i's belief about their own ability. That is:*

$$\begin{aligned} \partial^2\mu_j/\partial c_j\partial f_j &> 0 \ \forall\theta \ \text{if} \ \omega > 0 \\ \partial^2\mu_j/\partial c_j\partial f_j &< 0 \ \forall\theta \ \text{if} \ \omega < 0 \end{aligned} \tag{4}$$

In particular, if they believe women and men have the same ability, then j's gender does not matter. That is:

$$\partial^2\mu_j/\partial c_j\partial f_j = 0 \ \forall\theta \ \text{if} \ \omega = 0 \tag{5}$$

I consider deviations from these propositions are evidence of non-rationality.

5 Response to corrections

In this section, I document evidence that participants – both women and men – understand the notion of good and bad moves. Thus, they are more willing to collaborate with someone who contributed more to the puzzle. This is in line with what one would expect and validates my experimental design.

However, after controlling for that person's contribution, participants are less willing to collaborate with a person who corrected their move, even if that person makes good corrections, which is inefficient and seems to indicate deviation from the rational agent's benchmark in Proposition 1. However, it is unclear whether it is evidence of deviation from rationality; I will return to this point later in section 6.

Estimating equation I estimate the following model with OLS.

$$Select_{ij} = \beta_1 CorrectedGood_{ij} + \beta_2 CorrectedBad_{ij} + \beta_3 Female_j + \delta Contribution_j + \mu_i + \epsilon_{ij} \tag{6}$$

where each variable is defined as follows:

- $Select_{ij} \in \{0, 1\}$: an indicator variable equals 1 if i selects j as their collaborator, 0 otherwise.
- $CorrectedGood_{ij} \in \{0, 1\}$: an indicator variable equals 1 if j corrected i and moved the puzzle closer to the solution, 0 otherwise.
- $CorrectedBad_{ij} \in \{0, 1\}$: an indicator variable equals 1 if j corrected i and moved the puzzle far away from the solution, 0 otherwise.
- $Female_j \in \{0, 1\}$: an indicator variable equals 1 if j is female, 0 otherwise.
- $Contribution_j \in \mathbb{Z}$: js 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 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.[25]

The key identification assumption is that $Contribution_j$ fully captures j's ability perceived by i through j's puzzle moves (not true ability). This assumption is reasonable if we think participants' willingness to collaborate is increasing in the partner's contribution to the puzzle, which is consistent with that participants can partially observe their partners' ability and their expected utility is increasing in their payoff.

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

Results Table 3 presents the regression results of equation 6. Columns 1-4 include all participants' willingness to collaborate but column 1 excludes partner's contribution and individual fixed effects and column 2 partner's contribution. Column 3 combines good and bad correction as a single dummy variable. Columns 5-7 present the corresponding results for women and columns 8-10 for men.

Column 1 shows that when we do not control for between-participants variation, the coefficient estimate on good correction is underestimated. Column 2 shows that when we do not control for partner's contribution, the coefficient estimate on bad correction is negative and very large: the

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

Table 3: Response to corrections

Dependent variable:	Willing to collaborate (yes=1, no=0)									
Sample:	All					Female			Male	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Good correction	-0.208***	-0.238***		-0.204***	-0.269***		-0.229***	-0.197***		-0.168***
	(0.028)	(0.030)		(0.024)	(0.043)		(0.033)	(0.040)		(0.036)
Bad correction	-0.518***	-0.508***		-0.100***	-0.550***		-0.172***	-0.457***		-0.011
	(0.031)	(0.034)		(0.036)	(0.044)		(0.047)	(0.050)		(0.052)
Any correction			-0.198***			-0.237***			-0.152***	
			(0.022)			(0.030)			(0.031)	
Female partner	-0.003	-0.001	0.008	0.009	-0.009	0.002	0.004	0.007	0.016	0.016
	(0.016)	(0.017)	(0.014)	(0.014)	(0.021)	(0.018)	(0.018)	(0.026)	(0.021)	(0.021)
Partner's contribution			0.083***	0.084***		0.090***	0.089***		0.077***	0.080***
			(0.003)	(0.003)		(0.004)	(0.004)		(0.003)	(0.004)
Individual FE		✓	✓	✓	✓	✓	✓	✓	✓	✓
P-value: Good correction =Bad correction	0.000	0.000		0.020	0.000		0.347	0.000		0.016
Baseline mean	0.780	0.780	0.780	0.780	0.780	0.780	0.780	0.778	0.778	0.778
Baseline SD	0.414	0.414	0.414	0.414	0.414	0.414	0.414	0.416	0.416	0.416
Adj. R-squared	0.104	0.100	0.334	0.335	0.111	0.365	0.369	0.090	0.306	0.306
Observations	3180	3180	3180	3180	1670	1670	1670	1510	1510	1510
Individuals	464	464	464	464	244	244	244	220	220	220

Notes: This table presents the regression results of equation 6. Columns 1-4 include all participants' willingness to collaborate but column 1 excludes partner's contribution and individual fixed effects and column 2 partner's contribution. Column 3 combines good and bad correction as a single dummy variable. Columns 5-7 present the corresponding results for women and columns 8-10 for men. Baseline mean and standard deviation are willingness to collaborate with partners who do not make any corrections. Standard errors in parentheses are clustered at the individual level. Significance levels: * 10%, ** 5%, and *** 1%.

point estimate is -0.508 (p-value < 0.01); that is, participants are 50.8% less willing to collaborate with partners who made a bad correction, a correction that moved the puzzle far away from the solution. Indeed, these coefficient estimates are more negative than the coefficient estimates on good corrections: 0.271 more negative (p-value < 0.01). This is true when we separately examine women (column 5, 0.281 with p-value < 0.01) and men (column 8, 0.281 with p-value < 0.01).

Corroborating this, looking at column 3, the coefficient estimate on the partner's contribution is positive and quantitatively and statistically highly significant and is 0.083 (p-value < 0.01). This suggests that participants are 8.3% more willing to collaborate with partners who make one more good move. This is true for women (column 6, 0.090 with p-value < 0.01) and men (column 9, 0.077 with p-value < 0.01). This is evidence that my experimental design is valid: participants correctly understand the notion of good and bad moves and that participants are more willing to collaborate with partners who contributed more.

The coefficient estimate on any correction in column 3 is negative and quantitatively and statistically highly significant and is -0.198 (p-value < 0.01). This suggests that people are 19.8% less willing to collaborate with those who made a correction(s). To offset this effect, a partner's contribution has to increase by 0.79 standard deviation.[26] The corresponding coefficient estimate for women is -0.237 (column 6, p-value < 0.01) and -0.152 for men (column 9, p-value < 0.01).

26. The number is calculated as follows: $\hat{\beta}_{Partner's\ contribution} \times SD_{Partner's\ contribution} \times x = |\hat{\beta}_{Any\ correction}| \Rightarrow x = |\hat{\beta}_{Any\ correction}|/(\hat{\beta}_{Partner's\ contribution} \times SD_{Partner's\ contribution}) = 0.198/(0.09 \times 2.8) \approx 0.79$. $SD_{Partner's\ contribution} = 2.8$ is from panel B of Table 2 and is an arithmetic average of 2.73 for partners faced by women 2.87 for and partners faced by men: (2.73+2.87)/2=2.80.

Thus, participants are less willing to collaborate with a person who corrected their move.

This is not a problem if participants are more willing to collaborate with a person who made a good correction and less willing to collaborate with a person who made a bad correction. However, this is not the case: the coefficient estimate on good correction in column 4 is still negative and is -0.204 (p-value < 0.01). This suggests that people are less willing to collaborate even with those who made a good correction(s). The corresponding coefficient estimate for women is -0.229 (column 7, p-value < 0.01) and -0.168 for men (column 10, p-value < 0.01).

The coefficient estimate on bad correction in columns 4 is also negative and quantitatively and statistically significant and is -0.100 (p-value < 0.01). However, the magnitude is smaller than the coefficient estimate on good correction: the difference is -0.104 (p-value < 0.05). This is mainly driven by men: the corresponding coefficient estimate for women is -0.172 (column 6, p-value < 0.01) but is -0.011 (p-value > 0.10) for men.

These behaviors are inefficient. They also seem to indicate deviation from the rational agent's benchmark in Proposition 1. However, we cannot say anything definitive about rationality because response to corrections depends on the belief about people's own ability relative to partners and people are in general overconfident albeit that men are more overconfident (Croson and Gneezy 2009). Thus, these behaviors are not irrational if participants believe they are better at the puzzle than others. I will come back to this point in section 6.

6 Who respond negatively to corrections?

If only low ability people respond negatively to corrections, then correcting colleagues may not be very costly, because collaborations with those people will likely result in low quality outcomes. Also, it was inconclusive whether people's negative reaction to corrections documented in section 5 is irrational. However, if it is rational, people with higher puzzle solving ability – as measured in the part 1 individual practice stage – should respond less negatively to good corrections and more negatively to bad corrections because they are better able to distinguish good and bad corrections than people with lower puzzle solving ability as in Proposition 2.

In this section, I document that even high ability people respond negatively to good corrections with men responding more negatively, which indicates that (i) correcting colleagues is really costly both for individuals and groups and that (ii) negative response to corrections is irrational.

Estimating equation I estimate the following model with OLS.

$$\begin{aligned} Select_{ij} =& \beta_1 CorrectedGood_{ij} + \beta_2 CorrectedBad_{ij} + \beta_3 Female_j \\ &+ \beta_4 CorrectedGood_{ij} \times HighAbility_i + \beta_5 CorrectedBad_{ij} \times HighAbility_i \\ &+ \delta_1 Contribution_j + \delta_2 Contribution_j \times HighAbility_i + \mu_i + \epsilon_{ij} \end{aligned} \tag{7}$$

where each variable is defined as follows:

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

Other variables are as defined in equations 6.

Panel B of Figure 4 shows distribution of contribution of high ability people is indeed less dispersed, consistent with that I matched up high ability people with high ability people – and low ability people with low ability people – the collaborator selection stage in part 2.

Table 4: Response to corrections of high vs. low ability people

Dependent variable:	Willing to collaborate (yes=1, no=0)					
Sample:	All		Female		Male	
	(1)	(2)	(3)	(4)	(5)	(6)
Good correction		-0.155***		-0.208***		-0.107***
		(0.030)		(0.042)		(0.041)
Bad correction		-0.100**		-0.201***		0.005
		(0.047)		(0.064)		(0.063)
Any correction	-0.153***		-0.213***		-0.096**	
	(0.028)		(0.041)		(0.037)	
Female partner	0.008	0.009	0.002	0.002	0.015	0.014
	(0.014)	(0.014)	(0.018)	(0.018)	(0.021)	(0.021)
Partner's contribution	0.084***	0.084***	0.090***	0.089***	0.079***	0.082***
	(0.003)	(0.004)	(0.005)	(0.005)	(0.004)	(0.004)
Good correction x High ability		-0.118**		-0.048		-0.180**
		(0.050)		(0.066)		(0.075)
Bad correction x High ability		0.000		0.074		-0.061
		(0.072)		(0.095)		(0.109)
Any correction x High ability	-0.108**		-0.051		-0.152**	
	(0.044)		(0.061)		(0.064)	
Partner's contribution x High ability	-0.002	-0.001	-0.002	-0.001	-0.004	-0.003
	(0.005)	(0.005)	(0.007)	(0.007)	(0.007)	(0.008)
Individual FE	✓	✓	✓	✓	✓	✓
Baseline mean	0.780	0.780	0.783	0.783	0.778	0.778
Baseline SD	0.414	0.414	0.413	0.413	0.416	0.416
Adj. R-squared	0.335	0.336	0.365	0.368	0.308	0.308
Observations	3180	3180	1670	1670	1510	1510
Individuals	464	464	244	244	220	220

Notes: This table presents the regression results of equation 8. Columns 1-2 include all participants' willingness to collaborate. Columns 3-4 present the corresponding results for women and columns 5-6 for men. Baseline mean and standard deviation are willingness to collaborate with partners who do not make any corrections. Standard errors in parentheses are clustered at the individual level. Significance levels: * 10%, ** 5%, and *** 1%.

Results Table 4 presents the regression results of equation 8. As Table 3, columns 1-2 include all participants' willingness to collaborate. columns 3-4 the corresponding results for women and columns 5-6 for men.

In column 1, the coefficient estimate on the interaction between any correction and high ability is negative and statistically significant (p-value < 0.05). This effect mainly comes from men: the corresponding coefficient estimate for women (column 3) is less negative and statistically insignificant,

but is more for men (column 5, p-value < 0.05). Thus, high ability people, in particular men, dislike receiving corrections more than low ability people.

It is not a problem if this result is coming from high ability people responding less negatively or even positively to good corrections and more negatively to bad corrections. However, this is not the case: in column 2, the coefficient estimate on the interaction between good correction and high ability is negative (p-value < 0.05). This effect comes from both women and men, with the effect on men being stronger: the corresponding coefficient estimate for women (column 4) is negative albeit less so and statistically insignificant, and is more negative and statistically significant (p-value $<$ 0.05) for women (in column 6).

The coefficient estimate on the interaction between bad correction and high ability in column 2 is almost zero. The corresponding coefficient estimate is positive for women (column 4) and negative for men (column 6), although they are both statistically insignificant.

Thus, even high ability people respond negatively to good corrections with men responding more negatively. This suggests that correcting colleagues is really costly both for individuals and groups because even high ability people are less willing to collaborate with someone who corrects their mistakes.

This result also suggests that negative reaction to corrections is likely to be irrational: as discussed at the beginning of this section, high ability people should be able to distinguish good and bad corrections and should respond less negatively to good corrections and more negatively to bad corrections than low ability people as the rational agent benchmark in Proposition 2 suggests. However, what we see here is the opposite. At least, it is inconsistent.

7 Do women's corrections receive stronger negative reactions?

In this section, I document that people – either men or women – do not underestimate women's contribution, which suggests that their prior about women's ability to solve the puzzle is neither higher nor lower than men. However, I also document suggestive evidence that women's good corrections may receive stronger negative reactions by men, which is inconsistent with Proposition 3. This is a problem for women's career success because men are still majority in top positions in society and hence missing collaboration opportunities with them could be detrimental.

Estimating equation I estimate the following model with OLS.

$$\begin{aligned} Select_{ij} =& \beta_1 CorrectedGood_{ij} + \beta_2 CorrectedBad_{ij} + \beta_3 Female_j \\ &+ \beta_4 CorrectedGood_{ij} \times Female_j + \beta_5 CorrectedBad_{ij} \times Female_j \\ &+ \delta_1 Contribution_j + \delta_2 Contribution_j \times Female_j + \mu_i + \epsilon_{ij} \end{aligned} \tag{8}$$

Where each variable is defined as in equation 6.

Table 5: Response to corrections made by women vs. men

Dependent variable:	Willing to collaborate (yes=1, no=0)					
Sample:	All		Female		Male	
	(1)	(2)	(3)	(4)	(5)	(6)
Good correction		-0.187***		-0.248***		-0.104*
		(0.035)		(0.045)		(0.053)
Bad correction		-0.176***		-0.218***		-0.104
		(0.051)		(0.064)		(0.076)
Any correction	-0.203***		-0.260***		-0.125***	
	(0.031)		(0.042)		(0.045)	
Female partner	0.013	0.001	-0.001	-0.002	0.026	0.003
	(0.022)	(0.022)	(0.032)	(0.032)	(0.029)	(0.030)
Partner's contribution	0.084***	0.083***	0.090***	0.089***	0.078***	0.077***
	(0.004)	(0.004)	(0.006)	(0.006)	(0.005)	(0.006)
Good correction x Female partner		-0.035		0.035		-0.119*
		(0.044)		(0.057)		(0.067)
Bad correction x Female partner		0.144**		0.090		0.168
		(0.070)		(0.093)		(0.102)
Any correction x Female partner	0.009		0.047		-0.051	
	(0.041)		(0.056)		(0.059)	
Partner's contribution x Female partner	-0.002	0.002	-0.001	-0.001	-0.001	0.006
	(0.005)	(0.005)	(0.008)	(0.008)	(0.007)	(0.007)
Individual FE	✓	✓	✓	✓	✓	✓
Baseline mean	0.780	0.780	0.783	0.783	0.778	0.778
Baseline SD	0.414	0.414	0.413	0.413	0.416	0.416
Adj. R-squared	0.333	0.336	0.365	0.369	0.305	0.307
Observations	3180	3180	1670	1670	1510	1510
Individuals	464	464	244	244	220	220

Notes: This table presents the regression results of equation 8. Columns 1-2 include all participants' willingness to collaborate. Columns 3-4 present the corresponding results for women and columns 5-6 for men. Baseline mean and standard deviation are willingness to collaborate with partners who do not make any corrections. Standard errors in parentheses are clustered at the individual level. Significance levels: * 10%, ** 5%, and *** 1%.

Results Table 5 presents the regression results of equation 8. As Table 3, columns 1-2 include all participants' willingness to collaborate, columns 3-4 present the corresponding results for women and columns 5-6 for men.

Looking at column 1, the coefficient estimate on the interaction between partner's contribution and female partner is almost 0 and statistically insignificant. Column 3 shows this is true for women and column 5 for men. These suggest that people – both women and men – do not underestimate women's contribution when selecting a collaborator. In other words, people correctly believe that women and men are equally good at solving the puzzle.

In column 1, the coefficient estimates on the interaction between any correction and female partner is close to 0 and statistically insignificant. However, women and men respond differently: the corresponding coefficient estimate is positive for women (column 3) but negative for men (column 5), although they are statistically insignificant.

Column 2 splits any correction into good and bad correction and shows asymmetric response:

the coefficient estimates on the interaction between good correction and female partner is negative although statistically insignificant, but the coefficient estimates on the interaction between female partner and bad correction is positive (p-value < 0.05).

The negative coefficient estimate on the interaction between good correction and female partner mainly comes from men: looking at column 6, the corresponding coefficient estimate for men is -0.119 and marginally significant (p-value < 0.10), while for women it is 0.035 although statistically insignificant (column 4). On the other hand, the positive coefficient estimate on the interaction between female partner and bad correction comes from both women and men: the corresponding coefficient estimate is 0.090 for women (column 4) and 0.168 for men (column 6), although neither of them is statistically significant. Together with the evidence that men believe women are equally good at solving the puzzle as men, this is inconsistent with Proposition 3.

Men's less negative – or even positive – response to women's bad correction is a bit puzzling. One explanation is that men do not like to be corrected for their mistakes by women, but they are okay that women make mistakes.

Although only statistically marginally significant, men's negative reaction to women's good correction is worrying as men are still the majority in top positions both in industry (Bertrand 2018) and academia (Lundberg and Stearns 2019); missing collaboration opportunities with them could be detrimental for women's career.

8 Robustness checks

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

Excluding rounds 6 and 7 Remember that in rounds 6 and 7, participants' willing to collaborate is lower, they correct others more, and they are less likely to solve the puzzle, as shown in Figure 6 in section 3. As discussed in section 3, they are all outcomes of a particular pair independent of the type of the partner, but one may wonder whether these rounds are driving the results.

Excluding puzzles where good and bad corrections occurred There are 495 puzzles in which at least one correction occurred, of which 325 puzzles experienced good corrections only, 110 puzzles bad corrections only, and 60 puzzles experienced both good and bad corrections. In these 60 puzzles, it is unclear which corrections – good or bad – dominated people's mind in determining whether to collaborate with a paired person.

To address these concerns, I re-estimate equations 6, 7, and 8, and plot the coefficient estimates and 95% confidence intervals of the main coefficients of interest in Figures 7, 8, and 9, respectively, with solved puzzles only (green dots and lines), with rounds 1-5 only (red dots and lines), and with

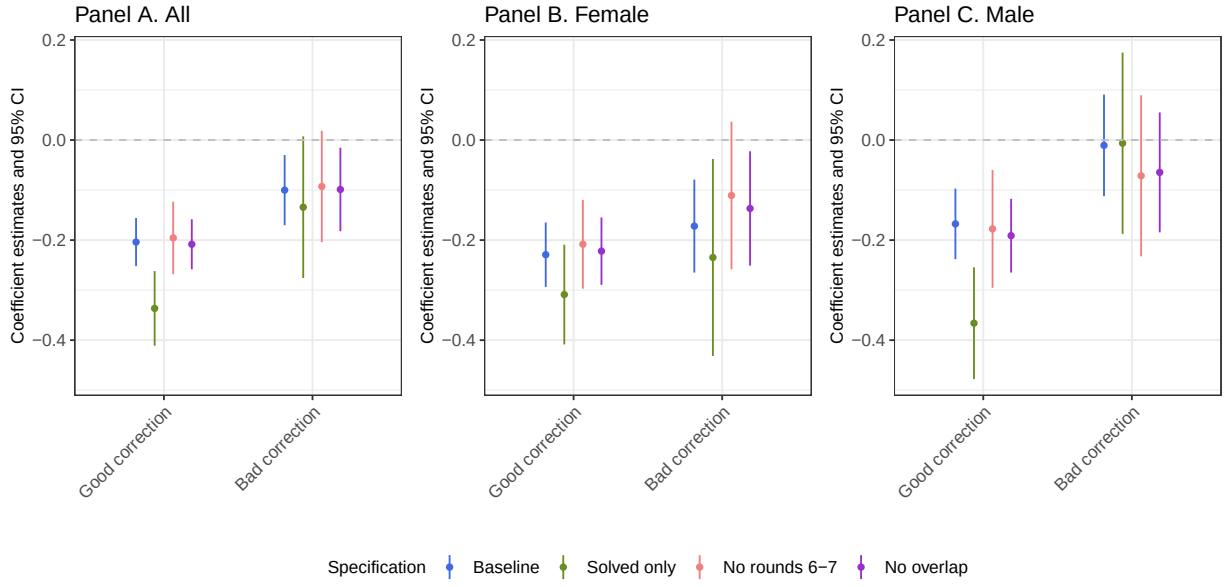
puzzles where only good or bad corrections occurred (purple dots and lines). As a reference, I also plot the coefficient estimates and 95% confidence intervals with the main sample used in Tables 3, 4, and 5 (blue dots and lines). All estimates are from the full models (columns 4, 7, and 10 for Table 3 and columns 2, 4, and 6 for Tables 4 and 5).

The main coefficients of interest for equation 6 are good and bad corrections. Looking at Figure 7, we see that most coefficient estimates are close to the main estimates. The estimates are more negative for good correction when the sample is limited to solved puzzles only, but they are more in line with the main findings.

The main coefficients of interest for equation 7 are the interaction between good correction and high ability and between bad correction and high ability. Looking at Figure 8, we again see most the coefficient estimates are close to the main estimates.

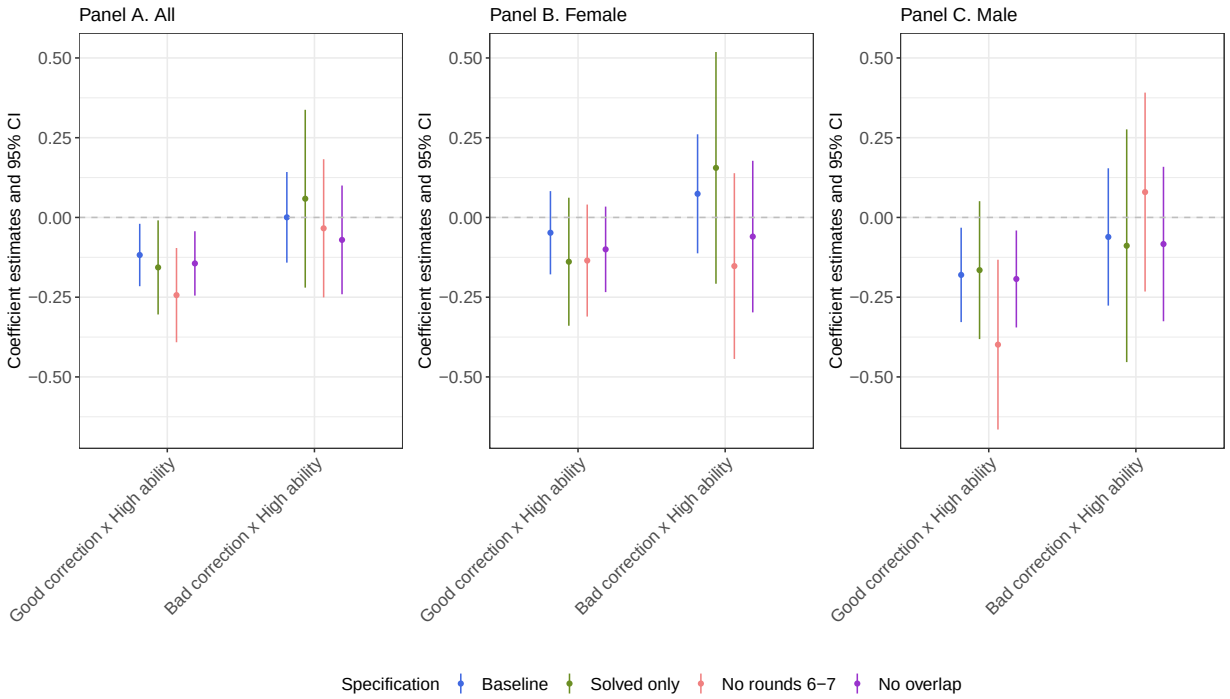
The main coefficients of interest for equation 8 are the interaction between good correction and female partner and between bad correction and female partner. Looking at Figure 9, we again see most the coefficient estimates are close to the main estimates. Again, the estimates with solved puzzles only present somewhat different evidence; in particular, response to good corrections by female partners is negative although statistically insignificant for women and positive for men. However, both estimates are very close to 0 and do not contradict with that the evidence that men react more negatively to women's good correction is only suggestive.

Figure 7: Response to corrections: Robustness



Notes: This figure plots the coefficient estimates and 95% confidence intervals of columns 4, 7, and 10 of Table 3 with solved puzzles only (the green dots and lines), with rounds 1-5 only (the red dots and lines), and with puzzles where only good or bad corrections occurred (the purple dots and lines). The blue dots and lines are the corresponding baseline estimates. They show that the findings in Table 3 are robust to limiting samples in these ways.

Figure 8: Response to corrections made of high vs. low ability people: Robustness



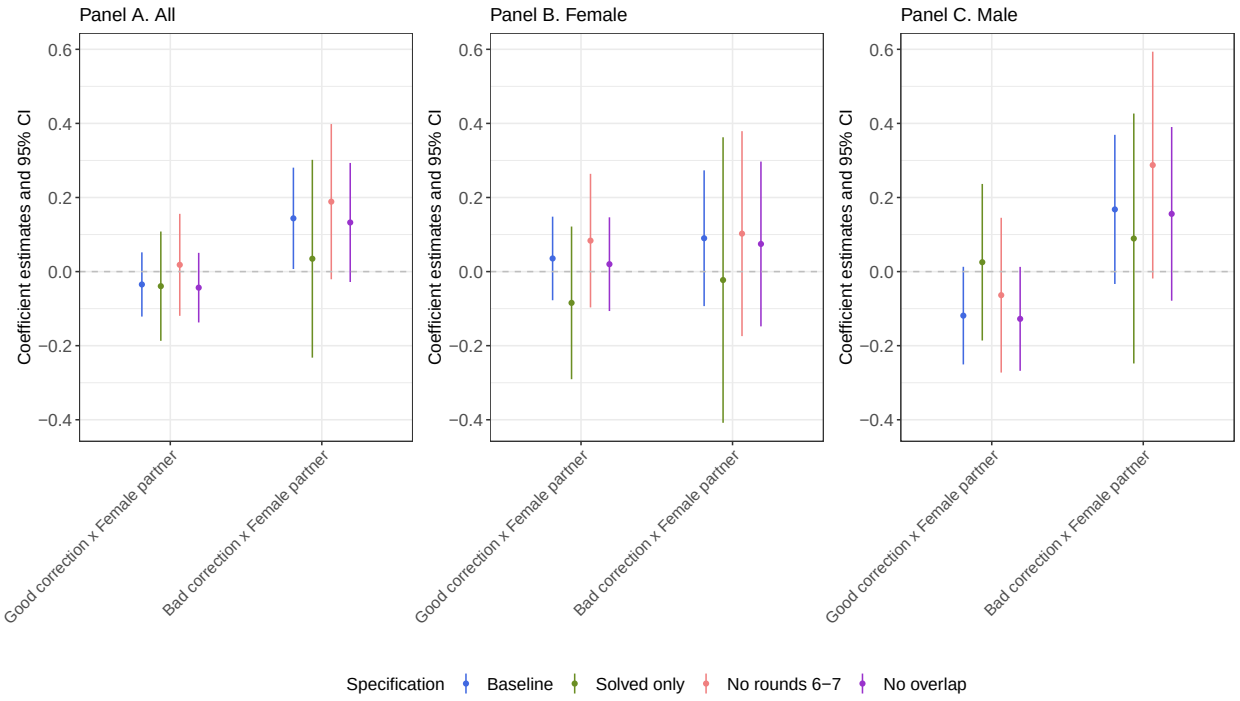
Notes: This figure plots the coefficient estimates and 95% confidence intervals of columns 2, 4, and 6 of Table 4 with solved puzzles only (the green dots and lines), with rounds 1-5 only (the red dots and lines), and with puzzles where only good or bad corrections occurred (the purple dots and lines). The blue dots and lines are the corresponding baseline estimates. They show that the findings in Table 4 are robust to limiting samples in these ways.

9 Conclusion

This paper demonstrates that people, including those with high productivity, are less willing to collaborate with a person who has corrected them even if the correction improves group performance. I also find suggestive evidence that men respond more negatively to women's corrections that improves group performance, presumably because men do not like to be corrected for their mistakes by women. Thus, dislike to be corrected distorts the optimal selection of talents and penalizes those who correct others' mistakes, and the distortion may be stronger when women correct men.

While a laboratory setting is different from the real-world, my findings are likely to be a lower bound because of the following three reasons. First, there is no reputation cost in my experiment: being corrected is not observed by others, unlike in the real-world. Second, the emotional stake is much smaller in my experiment: the puzzle ability is not informative of the ability relevant for their work or study – it is not something people have been devoting much of their time to, such as university exams, academic research, or corporate investment projects. Third, participants are equal in my experiment; in the real-world, there are sometimes senior-junior relationships, and corrections by junior people may induce stronger negative reactions. Thus, introducing reputation cost, using tasks that are more related to one's real-world ability, and having age variation would be interesting extensions of this paper.

Figure 9: Response to corrections made by women vs. men: Robustness



Notes: This figure plots the coefficient estimates and 95% confidence intervals of columns 2, 4, and 6 of Table 5 with solved puzzles only (the green dots and lines), with rounds 1-5 only (the red dots and lines), and with puzzles where only good or bad corrections occurred (the purple dots and lines). The blue dots and lines are the corresponding baseline estimates. They show that the findings in Table 5 are robust to limiting samples in these ways.

But there are two limitations. The first is that participants are strangers to each other in my experiments while people know each other in the real-world. Thus, it is possible that repeated interactions would mitigate people's negative response to corrections (but they may also magnify the negative response due to rivalry, failure to build a good rapport, etc.). The second limitation is that most participants are bachelor or master students who are supposed to have a weaker gender bias. The first point relates to the takeaway of my results: it would be worth investigating whether good workplace climate mitigates negative reaction to corrections. The second point relates to external validity: women's corrections may receive stronger and more robust negative reactions in the real-world where people are older.

Finally, my experiment is not designed to investigate the underlying mechanism, but the results are consistent with the findings of the literature on self-image concerns and information avoidance. For example, Kszegi (2006) finds that people avoid a difficult task when it reveals their ability. Corroborating this, Castagnetti and Schmacker (2021) find people select information less informative about their ability and Ewers and Zimmermann (2015) find people exaggerate their ability when others observe it even at the cost of reducing their payoff. There is also evidence that people respond differently to women's and men's feedback (Sinclair and Kunda 2000). In light of this literature, my results can be interpreted as follows: receiving good corrections is negative feedback and accepting

them as correct damages people's self-image.[27] Investigating how people's information avoidance interacts with workplace climate – and whether gender matters – would be an interesting future research.

27. Which means θ in the theoretical model in section 4 (equation 1) is not exogenous.

References

Abelson, Robert P. 1986. "Beliefs Are Like Possessions." *Journal for the Theory of Social Behaviour* 16 (3): 223–250.

Alan, Sule, Gozde Corekcioglu, and Matthias Sutter. 2021. *Improving Workplace Climate in Large Corporations: A Clustered Randomized Intervention.* Working Paper.

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

Bertrand, Marianne. 2018. "Coase Lecture – The Glass Ceiling." *Economica* 85 (338): 205–231.

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.

Castagnetti, Alessandro, and Renke Schmacker. 2021. *Protecting the Ego: Motivated Information Selection and Updating.* Working Paper.

Chen, Daniel L., Martin Schonger, and Chris Wickens. 2016. "oTree–An Open-Source Platform for Laboratory, Online, and Field Experiments." *Journal of Behavioral and Experimental Finance* 9:88–97.

Coffman, Katherine, Clio Bryant Flikkema, and Olga Shurchkov. 2021. "Gender Stereotypes in Deliberation and Team Decisions." *Games and Economic Behavior* 129:329–349.

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

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.

Edmans, Alex. 2011. "Does the Stock Market Fully Value Intangibles? Employee Satisfaction and Equity Prices." *Journal of Financial Economics* 101 (3): 621–640.

Egan, Mark, Gregor Matvos, and Amit Seru. 2021. "When Harry Fired Sally: The Double Standard in Punishing Misconduct." *Journal of Political Economy.*

Ewers, Mara, and Florian Zimmermann. 2015. "Image and Misreporting." *Journal of the European Economic Association* 13 (2): 363–380.

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* (US) 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.

Golman, Russell, David Hagmann, and George Loewenstein. 2017. "Information Avoidance." *Journal of Economic Literature* 55 (1): 96–135.

Greiner, Ben. 2015. "Subject Pool Recruitment Procedures: Organizing Experiments with ORSEE." *Journal of the Economic Science Association* 1 (1): 114–125.

Guiso, Luigi, Paola Sapienza, and Luigi Zingales. 2015. "The Value of Corporate Culture." *Journal of Financial Economics,* NBER Conference on the Causes and Consequences of Corporate Culture, 117 (1): 60–76.

Guo, Joyce, and María P. Recalde. 2022. "Overriding in Teams: The Role of Beliefs, Social Image, and Gender." *Management Science.*

Harrison, Glenn W., and John A. List. 2004. "Field Experiments." *Journal of Economic Literature* 42 (4): 1009–1055.

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

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

Kszegi, Botond. 2006. "Ego Utility, Overconfidence, and Task Choice." *Journal of the European Economic Association* 4 (4): 673–707.

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.

Lundberg, Shelly, and Jenna Stearns. 2019. "Women in Economics: Stalled Progress." *Journal of Economic Perspectives* 33 (1): 3–22.

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

Sinclair, Lisa, and Ziva Kunda. 2000. "Motivated Stereotyping of Women: Shes 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.

Wuchty, Stefan, Benjamin F. Jones, and Brian Uzzi. 2007. "The Increasing Dominance of Teams in Production of Knowledge." *Science* 316 (5827): 1036–1039. pmid: 17431139.

Appendix

Appendix A Advantages and drawbacks of quasi-laboratory format

The quasi-laboratory experimental format I use in this paper has advantages and drawbacks over the standard online and physical laboratory format.

Over standard online format The main advantage over standard online experimental format is that we can mostly avoid attrition, which is the main problem of online interactive experiments (Arechar, Gächter, and Molleman 2018). The reason is that compared to platforms such as MTurk and Prolific where participants' identity is fully anonymous by their 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. In my experiment, I experienced no participant attrition.

Another advantage is that we can fully control over who will participate in the experiment. For instance, we can screen out participants who have participated in particular kinds of experiments, such as experiments involving deception which is another problem of online experiments (Arechar, Gächter, and Molleman 2018). In my case, I have excluded participated who have participated in gender-related experiments in the past. This allows us to collect cleaner data.

The key drawback is the difficulty of collecting a large number of observations. Unlike MTurk or Prolific, the experimenter has to be present and respond to participants, if necessary, throughout the experiment. We could recruit a large number of participants at once, for example several hundreds, but it weakens the connection between the experimenter and the participants and can induce attrition.

Over physical laboratory format The main advantage over physical laboratory format is logistical convenience both for the experimenter and the participants: we can run and join experiments from our offices or home as long as we have a computer and an internet connection. It primarily benefited me to comply with the COVID-19 precautions. However, it also means that we can run laboratory experiments even if we do not have a physical laboratory in our university, for example in universities in low-income countries, as long as we set up ORSEE (Greiner 2015) or other subject management system, many of which are free.

Another advantage is that since participants can join the experiments from anywhere in the world, we can potentially run experiments with non-standard subjects or what Harrison and List (2004) call artefactual field experiments. For instance, non-student subjects or subjects in other countries. Although there can be regulation issues we have to overcome, it increases the kind of questions we can answer.

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

Appendix B Heterogeneity by gender bias

I have pre-specified in the pre-analysis plan that investigated heterogeneous response to women's corrections by the degree of gender bias measured by the six hostile and benevolent sexism questions used in Stoddard, Karpowitz, and Preece (2020), selected from Glick and Fiske (1996). Although I could not detect a meaningful variation, I present the results here.

I estimate the following model with OLS.

$$\begin{aligned} Select_{ij} =& \beta_1 CorrectedGood_{ij} + \beta_2 CorrectedBad_{ij} + \beta_3 Female_j \\ &+ \beta_4 CorrectedGood_{ij} \times HighBias_i + \beta_5 CorrectedBad_{ij} \times HighBias_i \\ &+ \delta_1 Contribution_j + \delta_2 Contribution_j \times HighBias_i + \mu_i + \epsilon_{ij} \end{aligned} \tag{B1}$$

where each variable is defined as follows:

- $HighBias_i \in \{0, 1\}$: an indicator variable equals 1 if i's gender bias score from the six hostile and benevolent sexism questions is above median among participants with the same gender (female or male), 0 otherwise.

Other variables are as defined in equations 6.

Table B1 presents the regression results of equation B1. As Table 3, columns 1-2 include all participants' willingness to collaborate. columns 3-4 the corresponding results for women and columns 5-6 for men.

In columns 1, 3, and 5, the coefficient estimate on the interaction among any correction, female partner, and high bias is negative but statistically insignificant. Also, in column 2, 4, and 6, the coefficient estimate on the interaction between good correction, female partner, and high bias as well as on the interaction between bad correction, female partner, and high bias are mostly negative but statistically insignificant. Thus, while the gender bias measure may detect some of the gender bias of participants, it is not variable enough to capture any meaningful heterogeneity among participants.

Appendix C Results with the original contribution measure

In the main text, I changed the definition of contribution from the one in the pre-analysis plan because there was truncation in the original contribution measure in more than 10% of the puzzle. Nonetheless, the same results hold when I use the original contribution measure, reported in Tables C1, C2, C3, and C4. Although the original measure is relative to one's pair while the measure I use in this paper is absolute, whether a measure is relative or absolute does not matter because I add individual fixed effects.

Table B1: Response to corrections made by women vs. men: Heterogeneity by gender bias

Dependent variable:	Willing to collaborate (yes=1, no=0)					
Sample:	All		Female		Male	
	(1)	(2)	(3)	(4)	(5)	(6)
Good correction		-0.180***		-0.260***		-0.066
		(0.046)		(0.058)		(0.071)
Bad correction		-0.196***		-0.185**		-0.201*
		(0.071)		(0.094)		(0.104)
Any correction	-0.215***		-0.267***		-0.137**	
	(0.041)		(0.053)		(0.065)	
Female partner	0.023	0.004	0.022	0.014	0.025	-0.006
	(0.029)	(0.029)	(0.039)	(0.040)	(0.041)	(0.042)
Partner's contribution	0.088***	0.087***	0.094***	0.094***	0.082***	0.079***
	(0.006)	(0.006)	(0.008)	(0.008)	(0.007)	(0.008)
Good correction x Female partner		-0.030		0.053		-0.151*
		(0.060)		(0.080)		(0.088)
Bad correction x Female partner		0.241**		0.179		0.312**
		(0.096)		(0.127)		(0.143)
Any correction x Female partner	0.036		0.086		-0.044	
	(0.055)		(0.076)		(0.079)	
Partner's contribution x Female partner	-0.005	0.001	-0.007	-0.004	-0.002	0.007
	(0.007)	(0.008)	(0.010)	(0.010)	(0.010)	(0.011)
Good correction x High bias		-0.018		0.027		-0.076
		(0.071)		(0.093)		(0.106)
Bad correction x High bias		0.036		-0.069		0.173
		(0.100)		(0.129)		(0.146)
Any correction x High bias	0.019		0.011		0.017	
	(0.063)		(0.089)		(0.091)	
Female partner x High bias	-0.027	-0.015	-0.050	-0.039	-0.010	0.008
	(0.044)	(0.044)	(0.064)	(0.065)	(0.058)	(0.059)
Partner's contribution x High bias	-0.009	-0.008	-0.008	-0.010	-0.010	-0.006
	(0.008)	(0.009)	(0.012)	(0.012)	(0.010)	(0.011)
Good correction x Female partner x High bias		-0.003		-0.025		0.060
		(0.089)		(0.118)		(0.134)
Bad correction x Female partner x High bias		-0.177		-0.170		-0.257
		(0.137)		(0.186)		(0.198)
Any correction x Female partner x High bias	-0.048		-0.073		-0.011	
	(0.082)		(0.114)		(0.118)	
Partner's contribution x Female partner x High bias	0.007	0.003	0.011	0.008	0.005	-0.000
	(0.011)	(0.011)	(0.016)	(0.016)	(0.014)	(0.014)
Individual FE	✓	✓	✓	✓	✓	✓
Baseline mean	0.781	0.781	0.783	0.783	0.779	0.779
Baseline SD	0.414	0.414	0.413	0.413	0.415	0.415
Adj. R-squared	0.333	0.335	0.363	0.368	0.304	0.305
Observations	3173	3173	1670	1670	1503	1503
Individuals	463	463	244	244	219	219

Notes: This table presents the regression results of equation B1. Columns 1-2 includes all participants willingness to collaborate. Columns 3-4 present the corresponding results for women and columns 5-6 for men. Baseline mean and standard deviation are willingness to collaborate with partners who do not make any corrections. Standard errors in parentheses are clustered at the individual level. Significance levels: * 10%, ** 5%, and *** 1%.

Table C1: Response to corrections: The original contribution measure

Dependent variable:	Willing to collaborate (yes=1, no=0)									
Sample:	All					Female			Male	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Good correction	-0.208***	-0.238***		-0.272***	-0.269***		-0.304***	-0.197***		-0.230***
	(0.028)	(0.030)		(0.026)	(0.043)		(0.035)	(0.040)		(0.038)
Bad correction	-0.518***	-0.508***		-0.160***	-0.550***		-0.234***	-0.457***		-0.065
	(0.031)	(0.034)		(0.037)	(0.044)		(0.048)	(0.050)		(0.054)
Any correction			-0.267***			-0.313***			-0.213***	
			(0.024)			(0.033)			(0.033)	
Female partner	-0.003	-0.001	0.006	0.008	-0.009	0.001	0.004	0.007	0.012	0.012
	(0.016)	(0.017)	(0.014)	(0.014)	(0.021)	(0.019)	(0.019)	(0.026)	(0.021)	(0.022)
Partner's contribution			1.181***	1.192***		1.171***	1.164***		1.196***	1.234***
			(0.054)	(0.058)		(0.076)	(0.078)		(0.076)	(0.084)
Individual FE		✓	✓	✓	✓	✓	✓	✓	✓	✓
P-value: Good correction =Bad correction	0.000	0.000		0.013	0.000		0.252	0.000		0.014
Baseline mean	0.780	0.780	0.780	0.780	0.780	0.780	0.780	0.778	0.778	0.778
Baseline SD	0.414	0.414	0.414	0.414	0.414	0.414	0.414	0.416	0.416	0.416
Adj. R-squared	0.104	0.100	0.309	0.314	0.111	0.320	0.330	0.090	0.300	0.300
Observations	3180	3180	3180	3180	1670	1670	1670	1510	1510	1510
Individuals	464	464	464	464	244	244	244	220	220	220

Notes: This table reports the same estimation results as Table 3 but with the original contribution measure specified in the pre-analysis plan, and show that the results are robust to using the original measure. Baseline mean and standard deviation are willingness to collaborate with partners who do not make any corrections. Standard errors in parentheses are clustered at the individual level. Significance levels: * 10%, ** 5%, and *** 1%.

Table C2: Response to corrections of high vs. low ability people: The original contribution measure

Dependent variable:	Willing to collaborate (yes=1, no=0)					
Sample:	All		Female		Male	
	(1)	(2)	(3)	(4)	(5)	(6)
Good correction		-0.216***		-0.269***		-0.168***
		(0.033)		(0.044)		(0.047)
Bad correction		-0.178***		-0.291***		-0.055
		(0.051)		(0.068)		(0.069)
Any correction	-0.224***		-0.284***		-0.167***	
	(0.031)		(0.045)		(0.042)	
Female partner	0.006	0.007	0.001	0.002	0.011	0.010
	(0.014)	(0.014)	(0.019)	(0.018)	(0.021)	(0.021)
Partner's contribution	1.200***	1.196***	1.208***	1.196***	1.195***	1.222***
	(0.068)	(0.073)	(0.101)	(0.101)	(0.089)	(0.098)
Good correction x High ability		-0.137***		-0.079		-0.192**
		(0.052)		(0.070)		(0.076)
Bad correction x High ability		0.049		0.144		-0.051
		(0.072)		(0.094)		(0.108)
Any correction x High ability	-0.101**		-0.060		-0.130**	
	(0.047)		(0.066)		(0.065)	
Partner's contribution x High ability	-0.039	-0.004	-0.088	-0.070	0.025	0.064
	(0.112)	(0.118)	(0.152)	(0.154)	(0.167)	(0.183)
Individual FE	✓	✓	✓	✓	✓	✓
Baseline mean	0.780	0.780	0.783	0.783	0.778	0.778
Baseline SD	0.414	0.414	0.413	0.413	0.416	0.416
Adj. R-squared	0.310	0.315	0.320	0.331	0.301	0.303
Observations	3180	3180	1670	1670	1510	1510
Individuals	464	464	244	244	220	220

Notes: This table reports the same estimation results as Table 4 but with the original contribution measure specified in the pre-analysis plan, and show that the results are robust to using the original measure. Baseline mean and standard deviation are willingness to collaborate with partners who do not make any corrections. Standard errors in parentheses are clustered at the individual level. Significance levels: * 10%, ** 5%, and *** 1%.

Table C3: Response to corrections made by women vs. men: The original contribution measure

Dependent variable:	Willing to collaborate (yes=1, no=0)					
Sample:	All		Female		Male	
	(1)	(2)	(3)	(4)	(5)	(6)
Good correction		-0.268***		-0.338***		-0.180***
		(0.038)		(0.049)		(0.058)
Bad correction		-0.232***		-0.275***		-0.167**
		(0.054)		(0.069)		(0.084)
Any correction	-0.277***		-0.348***		-0.192***	
	(0.034)		(0.049)		(0.048)	
Female partner	-0.053	-0.072	-0.099	-0.090	-0.011	-0.063
	(0.049)	(0.052)	(0.070)	(0.074)	(0.069)	(0.072)
Partner's contribution	1.115***	1.109***	1.064***	1.070***	1.159***	1.147***
	(0.082)	(0.085)	(0.116)	(0.114)	(0.116)	(0.125)
Good correction x Female partner		-0.008		0.063		-0.090
		(0.046)		(0.061)		(0.071)
Bad correction x Female partner		0.143*		0.085		0.188*
		(0.077)		(0.108)		(0.105)
Any correction x Female partner	0.023		0.069		-0.035	
	(0.044)		(0.063)		(0.062)	
Partner's contribution x Female partner	0.124	0.163	0.201	0.182	0.061	0.168
	(0.107)	(0.113)	(0.152)	(0.159)	(0.150)	(0.157)
Individual FE	✓	✓	✓	✓	✓	✓
Baseline mean	0.780	0.780	0.783	0.783	0.778	0.778
Baseline SD	0.414	0.414	0.413	0.413	0.416	0.416
Adj. R-squared	0.309	0.314	0.321	0.331	0.299	0.301
Observations	3180	3180	1670	1670	1510	1510
Individuals	464	464	244	244	220	220

Notes: This table reports the same estimation results as Table 5 but with the original contribution measure specified in the pre-analysis plan, and show that the results are robust to using the original measure. Baseline mean and standard deviation are willingness to collaborate with partners who do not make any corrections. Standard errors in parentheses are clustered at the individual level. Significance levels: * 10%, ** 5%, and *** 1%.

Table C4: Response to corrections made by women vs. men: Heterogeneity by gender bias, the original contribution measure

Dependent variable:	Willing to collaborate (yes=1, no=0)					
Sample:	All		Female		Male	
	(1)	(2)	(3)	(4)	(5)	(6)
Good correction		-0.254***		-0.327***		-0.151*
		(0.051)		(0.063)		(0.083)
Bad correction		-0.211***		-0.173*		-0.268***
		(0.067)		(0.091)		(0.096)
Any correction	-0.275***		-0.318***		-0.213***	
	(0.045)		(0.059)		(0.072)	
Female partner	0.020	-0.004	-0.006	0.005	0.041	-0.029
	(0.061)	(0.065)	(0.086)	(0.092)	(0.086)	(0.087)
Partner's contribution	1.254***	1.247***	1.246***	1.275***	1.251***	1.203***
	(0.097)	(0.103)	(0.138)	(0.139)	(0.137)	(0.149)
Good correction x Female partner		-0.047		0.025		-0.159
		(0.066)		(0.087)		(0.102)
Bad correction x Female partner		0.161*		0.042		0.327**
		(0.094)		(0.130)		(0.129)
Any correction x Female partner	0.001		0.046		-0.064	
	(0.059)		(0.080)		(0.089)	
Partner's contribution x Female partner	-0.020	0.033	0.021	0.007	-0.047	0.100
	(0.130)	(0.140)	(0.186)	(0.202)	(0.182)	(0.186)
Good correction x High bias		-0.037		-0.037		-0.054
		(0.076)		(0.100)		(0.117)
Bad correction x High bias		-0.040		-0.188		0.175
		(0.104)		(0.132)		(0.159)
Any correction x High bias	-0.014		-0.074		0.035	
	(0.069)		(0.100)		(0.098)	
Female partner x High bias	-0.166*	-0.152	-0.198	-0.191	-0.135	-0.095
	(0.099)	(0.104)	(0.139)	(0.143)	(0.139)	(0.145)
Partner's contribution x High bias	-0.303*	-0.299*	-0.382*	-0.404*	-0.212	-0.148
	(0.164)	(0.167)	(0.226)	(0.217)	(0.236)	(0.254)
Good correction x Female partner x High bias		0.085		0.089		0.134
		(0.094)		(0.125)		(0.141)
Bad correction x Female partner x High bias		-0.026		0.068		-0.225
		(0.151)		(0.212)		(0.200)
Any correction x Female partner x High bias	0.053		0.061		0.067	
	(0.089)		(0.128)		(0.124)	
Partner's contribution x Female partner x High bias	0.322	0.289	0.382	0.354	0.272	0.186
	(0.214)	(0.225)	(0.301)	(0.311)	(0.303)	(0.318)
Individual FE	✓	✓	✓	✓	✓	✓
Baseline mean	0.781	0.781	0.783	0.783	0.779	0.779
Baseline SD	0.414	0.414	0.413	0.413	0.415	0.415
Adj. R-squared	0.310	0.315	0.322	0.332	0.298	0.299
Observations	3173	3173	1670	1670	1503	1503
Individuals	463	463	244	244	219	219

Notes: his table reports the same estimation results as Table B1 but with the original contribution measure specified in the pre-analysis plan, and show that the results are robust to using the original measure. Baseline mean and standard deviation are willingness to collaborate with partners who do not make any corrections. Standard errors in parentheses are clustered at the individual level. Significance levels: * 10%, ** 5%, and *** 1%.

Appendix D Experimental instructions: English translation

App: pt0

Page: Reg

Registration

Please fill out the following information in order for us to pay you after the session. Please make sure that they correspond to the information you registered on ORSEE.

N.B. Please capitalize only the first letter of your first name and last name.

Good examples: Marco Rossi; Maria Bianchi; Anna Maria Gallo

Bad examples: MARCO ROSSI; maria bianchi; Anna maria Gallo

- First name: [Textbox]
- Last name: [Textbox]
- Email address registered on ORSEE: [Textbox]

[Check if there are any same first names. If so, add an integer (starting from 2) at the end of the first name]

Page: Draw

Draw a coin

Please draw a virtual coin by clicking the button below.

[Draw]

[Assign random number ranging from 1 to 40]

Page: Wait

Your coin

You drew the following coin.

Please wait until the session starts.

Page: Excess

Please click an appropriate button

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

Page: Intro

General instructions

Overview: This study will consist of **3 parts** and a follow-up survey and is expected to take **1 hour**. At the beginning of each part, you will receive specific instructions, followed by a set of understanding questions. You must answer these understanding questions correctly to proceed.

Your payment: For completing this study, you are guaranteed **2€** for your participation, but can earn up to **25€** depending on how good you are at the tasks. The tasks involve solving sliding puzzles, like the one shown below.

puzzle_2_0.png

Confidentiality: Other people participating in this study can see your first name. Aside from your first name, other participants will not see any information about you. **At the conclusion of the study, all identifying information will be removed and the data will be kept confidential**. If there is more than one participant with the same first name, we add a number at the end of your first name (e.g. Marco2).

General rules: During the study, please turn off your camera and microphone, and do not communicate with anyone other than us. Also, please do not reload the page or close your browser because it may make your puzzle unsolvable. If you have any questions or face any problems, please send us a private chat on Zoom.

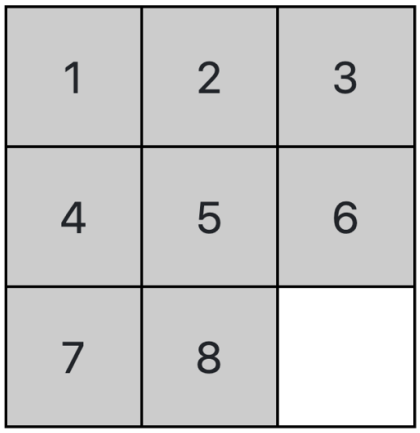
App: pt1

Page: Intro

Instructions for part 1 out of 3

In this part, you will solve the puzzle alone to familiarize yourself with it. You can solve as many puzzles as possible (but a maximum of 15 puzzles) in **4 minutes**. You will earn **0.2€ for each puzzle** you solve.

Your goal is to move the tiles and order them as follows:



puzzle_goal.png

Before you start, please go through the three examples below to understand how to solve the puzzle.

Example 1:

First, consider the following puzzle.

puzzle_1.png

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

Among the 3 moves, moving 6 up is the only correct move: by moving 6 up, you can solve the puzzle. The other moves do not solve the puzzle.

When you click a tile next to an empty cell, the tile will be moved to the empty cell. So, in this case, you should click 6 to move it up.

Example 2:

Next, consider the following puzzle.

puzzle_2_0.png

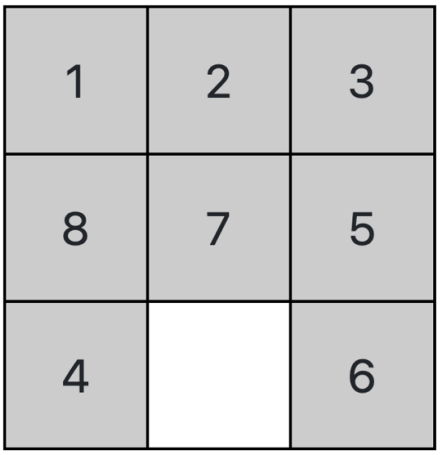
First, there are 2 moves you can make: move 2 right and move 3 up. Which moves should you make?

Observe that the only tiles that are not in the correct order are 3 and 6. So, you should move 3 up.

After moving 3 up, the puzzle will look like the one in example 1. Then you should move 6 up and the puzzle will be solved.

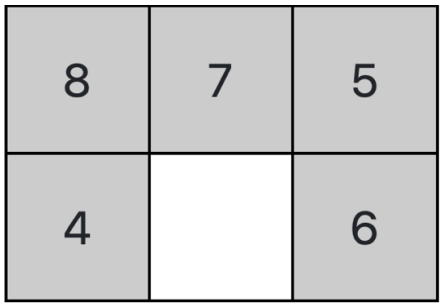
Example 3:

Finally, consider the following puzzle.



puzzle_3_0.png

This puzzle is a bit complicated but observe that the top row is already in the correct order. So, let's keep the top row as is, and think about the remaining part. **When the top row is in the correct order, you should always keep it as is**. So, think of this puzzle as the following simpler puzzle.



puzzle_3_0_2x3.png

You could solve the puzzle by trial and error. However, **after making the top row in the correct order, you should next make the left column in the correct order** to solve the puzzle faster. There are two moves you can make: move 4 right and move 7 down. Which is the faster way to make the left column in the correct order?

Let's try moving 4 right.

puzzle_3_1_bad_0.png

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

puzzle_3_1_bad_1.png

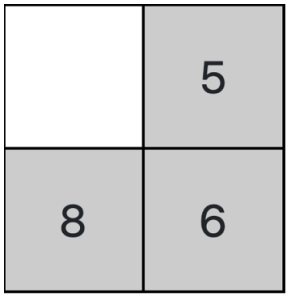
Now, if you ignore the top row which is already in the correct order, the only tile you can move is 7. So, let's move it to the left.

puzzle_3_1_bad_2.png

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

puzzle_3_1_bad_3.png

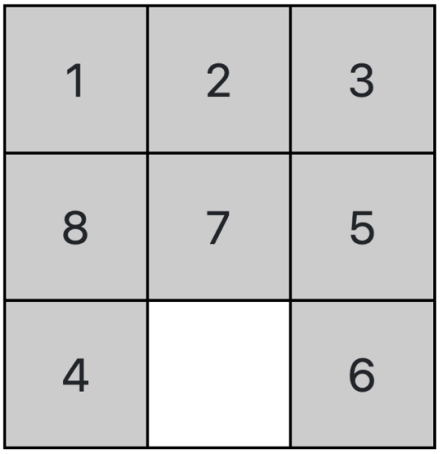
Now let's also keep the left column as is.



puzzle_3_1_bad_3_2x2.png

Then you can solve the puzzle by moving 5 left and then 6 up. With this method, **you have moved tiles nine times in total**.

Let's go back to the initial puzzle.



puzzle_3_0.png

This time, let's try moving 7 down.

puzzle_3_1_good.png

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

puzzle_3_4_good.png

Let's keep the left column as is (as well as the top row).

puzzle_3_4_good_2x2.png

Now it's easy to solve the puzzle: move 8 down, 5 left, and 6 up. With this method, **you have only moved tiles seven times in total**.

Because there is a time limit, it's better to solve the puzzle with the minimum number of moves. **We call a move a good move if it makes a puzzle closer to the solution, and a bad move if it makes a puzzle far from the solution. There are no neutral moves: all moves are either good or bad.**

In summary: when you solve the puzzle, first make the top row in the correct order, then make the left column in the correct order. Always try to make the number of moves as small as possible.

Understanding questions:

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

1. Which of the following statements is true?

- ✓In this part, I will work on the puzzles individually for 4 minutes and earn 0.2€ for each puzzle I solve.
- In this part, I will work on the puzzles in pairs for 4 minutes and earn 0.2€ for each puzzle we solve.
- In this part, I will work on the puzzles individually for 4 minutes, but I will not earn anything.

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

- A
- ✓B

A

puzzle_2_0.png

B

puzzle_goal.png

3. What is the strategy you should use to solve the puzzle as fast as possible?

- First, make the left column in the correct order, then the bottom row. Always minimize the number of moves I make.
- First, make the top row in the correct order, then the right column. Always minimize the number of moves I make.
- ✓First, make the top row in the correct order, then the left column. Always minimize the number of moves I make.

4. Look at the following puzzle. Which is the good move?

- Move 4 down.
- ✓Move 7 left.

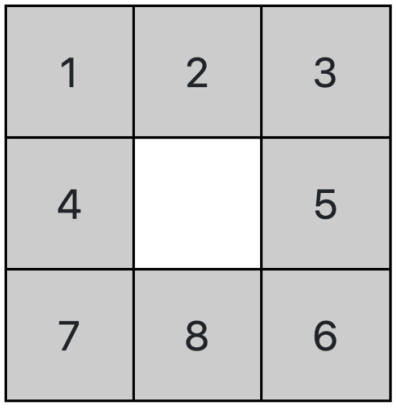
puzzle_3_3_good.png

5. Consider the puzzle in question 4. What is the minimum number of moves to solve the puzzle?

- 2
- 3
- ✓4

6. Look at the following puzzle. Which is the good move?

- ✓Move 5 left.
- Move 8 up.



puzzle_3_5_good.png

7. Consider the puzzle in question 6. What is the minimum number of moves to solve the puzzle?

- ✓2
- 3
- 4

Page: Ready

Be ready

[5 seconds time count]

Please be ready for the individual round.

Page: Game

Individual round

[4 minutes time count]

[max. 15 puzzles with increasing difficulty]

Page: Proceed

The individual round is over

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

Please click Next to proceed.

App: pt2

Page: Intro

Instructions for part 2 out of 3

In this part, you will **choose your partner for part 3**, the next part.

Although you will not earn anything in this part, it is important to choose the best partner possible: in part 3, you will work on the puzzles for 12 minutes in a pair by moving the tiles in turn, and both you and your partner will earn 1€ for each puzzle you two solve. There is a maximum of 20 puzzles you and your partner can solve (so the maximum earning is 20€).

You will **meet 7 other people** participating in this session one by one and solve 1 puzzle together by moving tiles in turn as you would do in part 3. One of you will be randomly chosen to make the first move at the beginning of each puzzle. You will have a **2-minute limit** for each puzzle.

After solving the puzzle, you will **choose whether you want to work with this person in part 3 too**. This person or other people in this session will not see your choice. **You can choose as many people as you want**.

After you meet all the 7 people and state your choices, we will check all the choices you and the 7 other people have made, and decide each person's partner for part 3 as follows:

1. We randomly choose 1 person out of you and the other 7 people. Call this person Giovanni.
2. We then check if Giovanni has a "match": among people Giovanni has chosen, we check whether these people also have chosen Giovanni. If there is such a person, we make Giovanni and this person as partners for part 3.
3. If Giovanni has more than one match, we randomly choose one of the matches and make them as partners for part 3.
4. If Giovanni has not chosen anyone, the people Giovanni has chosen have not chosen Giovanni, or those people already have their partner, we put Giovanni on a waiting list and repeat points 1-3 above.
5. After we choose all people, we randomly match people on the waiting list as partners for part 3.

So, **even if you choose a particular person, you may not be able to work with that person in part 3**. So, choose everyone whom you want to work with in part 3.

<u>Understanding questions</u>:

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

1. Which of the following statements is true?

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

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

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

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

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

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

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

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

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

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

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

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

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

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

- Giovanni
- Valeria

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

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

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

Page: Puzzle

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

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

[2 minutes time count]

Page: Pref

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

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

[Yes, No]

App: pt3

Page: Partner

Your partner for part 3

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

Page: Intro

Instructions for part 3 out of 3

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

Understanding questions:

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

1. Which of the following statements is true?

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

2. You and your partner…

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

Page: Ready

Be ready

[5 seconds time count]

Please be ready for the group round.

Page: Game

Puzzle 1/2/3/…/20

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

[12 minutes time count]

[max. 20 puzzles with increasing difficulty]

Page: Proceed

The group round is over

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

Please click Next to proceed.

App: pt4

Page: Intro

A follow-up survey

As the last task, we will ask you a series of questions in which there are no right or wrong answers. We are only interested in your personal opinions. We are interested in what

characteristics are associated with people's behaviors in this study. **The answers you provide will in no way affect your earnings in this study and are kept confidential.**

Please click Next to start the survey.

Page: SurveyASI

Survey page 1 out of 2

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

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

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

Page: SurveyDem

Survey page 2 out of 2

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

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

Page: ThankYou

Thank you for your participation

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

Your earnings:

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

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

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

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

App: pt99

Page: ThankYou

Thank you for showing up

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

Chapter 2

The Role of Gender and Cognitive Skills on Other People's Generosity

Yuki Takahashi*

Abstract

Cognitive skills are an important personal attribute that affects career success. However, colleagues' support is also vital as most works are done in groups, and the degree of their support is influenced by their generosity. Social norms enter in groups, and gender may interact with cognitive skills through gender norms in society. Because these gender norms penalize women with high potential, they can reduce colleagues' generosity towards these women. Using a novel experimental design where I exogenously vary gender and cognitive skills and sufficiently powered analysis, I find neither the two attributes nor their interactions affect other people's generosity; if anything, people are more generous to women with high potential. I argue that my findings have implications for the role of gender norms in labor markets.

JEL Classification: J16, M54, D91, C91
Keywords: gender, cognitive skills, social norm, generosity, dictator game

*Department of Economics, University of Bologna. Email: yuki.takahashi2@unibo.it. I am grateful to Maria Bigoni, Siri Isaksson, and Natalia Montinari whose feedback was essential for this project, and to participants of the experiment for their participation and cooperation. Laura Anderlucci, Tiziano Arduini, Francesca Barigozzi, Enrico Cantoni, Chiara Natalie Focacci, Margherita Fort, Catalina Franco, Fabio Landini, Annalisa Loviglio, Valeria Maggian, Joshua Miller, Monika Pompeo, Eugenio Proto, Tommaso Sonno, Alessandro Tavoni, Bertil Tungodden, ESA Experimental Methods Discussion group, and the University of Bologna's PhD students all provided many helpful comments. This paper also benefited from participants' comments at the Applied Young Economist Webinar, the BEEN Meeting, seminars at Ca' Foscari University, the NHH, and the University of Bologna. Veronica Rattini and oTree help & discussion group kindly answered my questions about oTree programming. Lorenzo Golinelli provided excellent technical and administrative assistance. This study was pre-registered with the OSF registry (https://osf.io/r6d8f/files).

1 Introduction

Cognitive skills are an important personal attribute that affects career success (Herrnstein and Murray 1996).[1] However, colleagues' support is also vital because most works are done in groups (Jones 2021; Lazear and Shaw 2007; Wuchty, Jones, and Uzzi 2007) and the degree of their support is influenced by their generosity. Social norms enter in groups, and gender may interact with cognitive skills through gender norms in society. Because these gender norms penalize women with high potential as those women are inconsistent with the stereotypical women (Eagly and Karau 2002; Heilman 2001; Ridgeway 2001; Rudman and Phelan 2008), they can reduce colleagues' generosity towards these women.[2]

This paper studies how gender, cognitive skills, and their interaction affect other people's generosity, focusing on women with high cognitive skills. Answering this question using secondary data is difficult due to non-random group formation and that cognitive skills are correlated with economic preferences and hence with generosity (Falk et al. 2021). Also, a clean measure of other people's generosity is not readily available in secondary data.

Thus, I design a laboratory experiment where participants first work on an incentivized IQ test which measures cognitive skills. After the test, participants are randomly assigned to a group of six and receive a ranking of their IQ within their group. Then three of the six members are randomly chosen to be dictators and play three rounds of dictator game with the other three members chosen to be recipients, observing the recipients' facial photos and first names, both of which convey information about gender, and the IQ ranks. The dictators' allocation is used as a measure of generosity. The use of photos follows recent literature and allows the dictators to infer the gender of the recipients naturally as they would do in the real world (Babcock et al. 2017; Coffman 2014; Isaksson 2018). I use dictator IQ fixed effects in the analysis to compare allocations of dictators with the same cognitive skills but assigned different IQ ranks due to random group formation.

I find neither gender, IQ, nor their interactions affect dictators' allocation: the point estimate is quantitatively negligible and statistically indistinguishable from 0, and the confidence interval is tight; if anything, women with higher IQ receive more allocation. This result is not driven by the so-called "beauty premium." The results hold across the whole distribution and even when I separately examine female and male dictators' allocation. Although statistically insignificant, belief about paired recipients' IQ is roughly consistent with the experimental design. These findings suggest that one's gender, cognitive skills, or their interaction do not play a significant role in other people's generosity.

This paper primarily relates to studies on the role of gender norms in one's career. The literature finds that people perceive female leaders (Heilman, Block, and Martell 1995; Heilman and Okimoto 2007; Rudman and Kilianski 2000) and competent women (Heilman et al. 2004; Rudman 1998)

1. Yet another prominent attribute is non-cognitive skills: (Cawley, Heckman, and Vytlacil 2001; Cunha and Heckman 2008; Heckman, Stixrud, and Urzua 2006).

2. Indeed, gender affects one's career through structural problems in labor markets such as unequal burden of family and child care (Bertrand 2018; Goldin 2014) and labor market norms designed for men who are more risk-loving and like competition (Bertrand 2011; Croson and Gneezy 2009; Dohmen et al. 2011; Niederle and Vesterlund 2011).

negatively.[3][4] Evidence from laboratory experiments shows that female leaders (Chakraborty and Serra 2021) and competitors (Datta Gupta, Poulsen, and Villeval 2013) receive more aggressive treatments and less support from men (Born, Ranehill, and Sandberg 2020). Nevertheless, evidence from audit studies is mixed: while Quadlin (2018) finds top-performing female college students receive less favorable treatment in hiring than equally qualified male students, Ceci and Williams (2015) and Williams and Ceci (2015) find qualified female candidates for assistant professors receive equal or more favorable treatment than equally qualified male candidates.[5] Also, Bursztyn, Fujiwara, and Pallais (2017) find that unmarried female MBA students behave in a less career-ambitious way in front of male classmates. My results suggest that these studies' findings are not likely to be driven by violation of cognitive skill-related gender norms.

This paper also contributes to the literature on the role of gender in dictator games. Bolton and Katok (1995) and Boschini, Muren, and Persson (2012) find that female and male dictators allocate the same amount, while Chowdhury, Grossman, and Jeon (2019), Dreber et al. (2013), and Eckel and Grossman (1998) find that female dictators allocate more. Bilén, Dreber, and Johannesson (2021) find that although female dictators allocate more, it is not quantitatively significant. Andreoni and Vesterlund (2001) find that the role of a dictator's gender on allocation depends on the price of allocation: female dictators allocate more when doing so reduces their own earnings while male dictators allocate more when doing so does not reduce their own earnings so much. Klinowski (2018) finds that female dictators allocate so that the amount between themselves and recipients are equalized, but aside from that, female and male dictators allocate the same amount. Aguiar et al. (2009) find that people expect female dictators to allocate more. Rosenblat (2008) finds that female dictators allocate more to physically attractive women and men than male dictators. Aksoy, Chadd, and Koh (2021) find that Republican heterosexual people allocate less to LGBTQ+ people. My paper enriches this literature by introducing cognitive skills in the role of gender in dictator game allocation.

The remainder of the paper proceeds as follows. Section 2 describes the experimental design, procedure, and implementation. Section 3 describes the data. Section 4 presents the main results. Section 5 show robustness of the main results. Section 6 concludes.

3. The literature also finds that people even evaluate competent women negatively, but these results are obtained in set-ups without real consequences (Phelan, MossRacusin, and Rudman 2008; Rudman and Fairchild 2004; Rudman et al. 2012).

4. The literature also finds that people penalize male losers (Cappelen, Falch, and Tungodden 2019; Moss-Racusin, Phelan, and Rudman 2010) and LGBTQ+ people (Aksoy, Chadd, and Koh 2021; Gorsuch 2019). These are equally important issues which we have to deal with.

5. However, Håkansson (2021) finds that female politicians, especially those in high positions, receive unfavorable treatment using Swedish data.

2 Experiment

2.1 Design and procedure

The experiment consists of two parts. Participants receive instructions at the beginning of each part. They earn a participation fee of 2.5€ for their participation.

Pre-experiment: Random desk assignment & photo taking

After registration at the laboratory entrance, participants are randomly assigned to a partitioned computer desk. Afterwards, participants have their facial photos taken at a photo booth and enter their first name on their computer. After that, the experimenters go to each participant's desk to check that their photo and first name match them to ensure all participants that other participants' photos and first names are real, following Isaksson (2018).

Part 1: IQ test

In part 1, participants work on an incentivized 9 IQ questions for 9 minutes. I use Bilker et al. (2012)'s form A 9-item Raven test which measures one's IQ more than 90% as good as the full-length Raven test. Participants receive 0.5€ for each correct answer, and they do not receive information about how many IQ questions they have solved correctly until the end of the experiment.

After the IQ test, participants make an incentivized guess on the number of IQ questions they have solved correctly; they receive 0.5€ if their guess is correct. The answer to this question measures their over-confidence level. They do not receive feedback on their guess until the end of the experiment.

Following Eil and Rao (2011), six participants are randomly grouped and informed of the ranking of their IQ relative to other group members. Ties are broken randomly. They then answer a set of comprehension questions about their IQ rank; they cannot proceed to the next part until they answer these questions correctly.

Part 2a: Dictator game (dictators only)

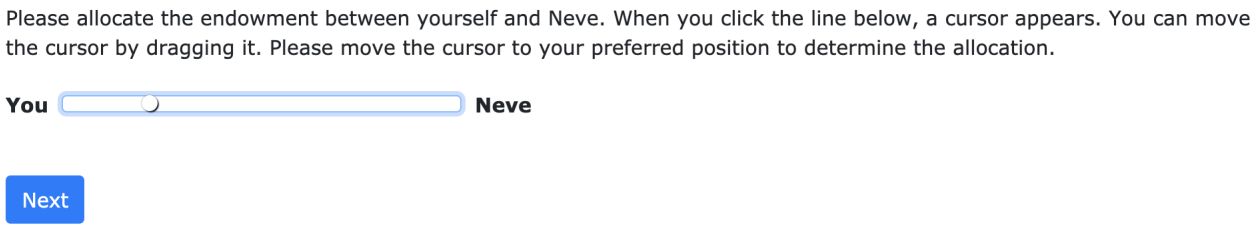
In part 2, three participants in each group are randomly assigned to the role of dictators and the other three participants the role of recipients. Dictators are paired with the three recipients in their group one by one in a random order, receive an endowment, and play a dictator game. Thus, they play a dictator game three times with three different recipients. When they play the dictator game, dictators observe the recipients' facial photo and first name and IQ rank; see panel A of figure 1 for an example. The use of photos follows recent literature (Babcock et al. 2017; Coffman 2014; Isaksson 2018) and minimizes experimenter demand effects. While I use photos, I show later that the results are not driven by the so-called "beauty premium."

Dictators are also told that their allocation decisions are anonymous: they are told that their allocation will be paid to the recipients as a "top-up" to their earnings. Dictators decide allocation

Figure 1: Dictator's allocation screen

(a) Initial screen

(b) After clicking the slider



Notes: This figure shows an example of a dictator's allocation screen. Panel A shows the screen before clicking the slider bar and panel B after clicking it. In this example, the dictator is playing the first round and paired with a recipient whose first name is Neve with IQ rank 5.

by moving a cursor on a slider where the cursor is initially hidden to prevent anchoring; panel B of figure 1 shows the cursor after clicking the slider. I vary the endowment across rounds to make each dictator game less repetitive: 7€ for 1st and 3rd rounds, 5€ for 2nd round. At the end of the experiment, one out of three allocations is randomly chosen for each participant as earnings for this

part.[6]

Part 2b: Belief elicitation (recipients only)

I also collect an indirect measure of dictators' beliefs on how many IQ questions the paired recipients have solved correctly. To prevent the belief elicitation to affect or be affected by the dictator game, I exploit the random assignment of participants to dictators and recipients (derived from the random desk assignment) and use recipients' beliefs as a proxy for dictators' beliefs. Specifically, while dictators are playing the dictator game, recipients are paired with the other two recipients in the same group one by one in random order and make incentivized guesses on how many IQ questions they have solved correctly, observing the other two recipients' facial photo, first name, and IQ rank. Each correct guess gives them 0.5€.

To address the non-anonymity of showing facial photos and first names, I ask participants how well they know the paired participants on a scale of 4.[7] I ask this question twice to make sure they do not answer randomly: right after the three dictator games for dictators or two guesses for recipients and in the post-experimental questionnaire.

Post-experiment: Questionnaire

After the dictator game and guessing are over, participants are told their earnings from the IQ test, dictator game, and the guesses. Before receiving their earnings, participants answer a short questionnaire about their demographics that are used for balance tests and robustness checks. Recipients are also asked if I could use their photo in another experiment with a gratuity of 1.5€.

2.2 Implementation

The experiment was programmed with oTree (Chen, Schonger, and Wickens 2016) and conducted in English during November-December 2019 at the Bologna Laboratory for Experiments in Social Science (BLESS). I recruited 390 students (195 female and 195 male) of the University of Bologna via ORSEE (Greiner 2015) who (i) were born in Italy, (ii) had not participated in gender-related experiments in the past (as far as I know), and (iii) available to participate in English experiments. The first condition is to reduce the chance that recipients' first names and photos signal ethnicity, race, or cultural background. The second condition is to reduce experimenter demand effects. The third condition is to run the experiment in English. The number of participants was based on the power simulation in the pre-analysis plan to achieve 80% power.[8] The experiment is pre-registered with the OSF.[9]

6. For each dictator for each round, one of the three recipients in the same group is randomly chosen *without replacement* and the dictator allocates the endowment between themselves and the recipient. Thus, it is possible that two dictators play dictator game with the same recipient in the same round. At the end of the dictator games, each participant has three allocations, and one of which is randomly chosen for payment.

7. The answer choices are: "I didn't know him/her at all," "I saw him/her before," "I knew him/her but not very well," and "I knew him/her very well."

8. I exclude the 1st session data because of the problem discussed in appendix A.

9. The pre-registration documents are available at the OSF registry: https://osf.io/r6d8f/files.

As a further attempt to make the data cleaner, I exclude recipients with non-Italian sounding names and allocations in which the dictator declared they knew the paired recipients "very well" at least once.[10] These data screenings leave me 388 participants, 195 dictators, and 558 dictators' allocations.

I ran 24 sessions in total, and the number of participants in each session was a multiple of 6 (12 to 30). The average length of a session was 70 minutes, including registration and payment. The average payment per participant was about 10€ including the participation fee and 1.5€ of gratuity for photo use in another experiment (only for those recipients who agreed).

3 Data description

Table 1 describes own (panel A) and paired participants' characteristics (panel B) as well as dictators' social distance with paired recipients (panel C) and dictator game allocation (panel D).

Panel A shows that female dictators solve 0.37 fewer IQ questions (out of 9) than male dictators, but the difference is quantitatively insignificant. Also, female dictators are more likely to major in humanities and less likely to major in STEM fields, consistent with a pattern observed in most OECD countries (see, for example, Carrell, Page, and West 2010). In addition, female dictators are less overconfident than male dictators, another pattern observed in other studies (Bertrand 2011; Croson and Gneezy 2009; Niederle and Vesterlund 2011). Further, women are more likely to have finished undergraduate studies, consistent with that women are more educated than men in OECD countries (see, for example, Almås et al. 2016; Autor and Wasserman 2013).

Panel B shows that paired participants' characteristics are roughly balanced, except that female dictators are 10% more likely to be paired with recipients from the Emilia-Romagna region where the experiment was conducted.

Panel C shows that dictators do not know about 95-98% of the paired recipients, mitigating the concern that dictator game allocation is driven by social distance outside the laboratory. To elaborate on this point, Figure 2 plots empirical CDF of dictators' allocation, which resembles that of Bohnet and Frey (1999)'s one-way identification with information treatment where the social distance between dictators and recipients is the closest to my setting.

Panel D shows that female dictators allocate their endowment to paired recipients 6% more than male dictators, although the difference is only marginally significant at 10%. This observation is consistent with a meta-analysis that women give more, but the difference is not quantitatively large (Bilén, Dreber, and Johannesson 2021). Residualized dictator game allocation shows the allocation after adding the dictator IQ fixed effects, my empirical approach to address the endogeneity of dictators' cognitive skills in the analysis explained later, still has enough variation, suggesting that the dictator IQ fixed effects do not over-control dictator game allocation.

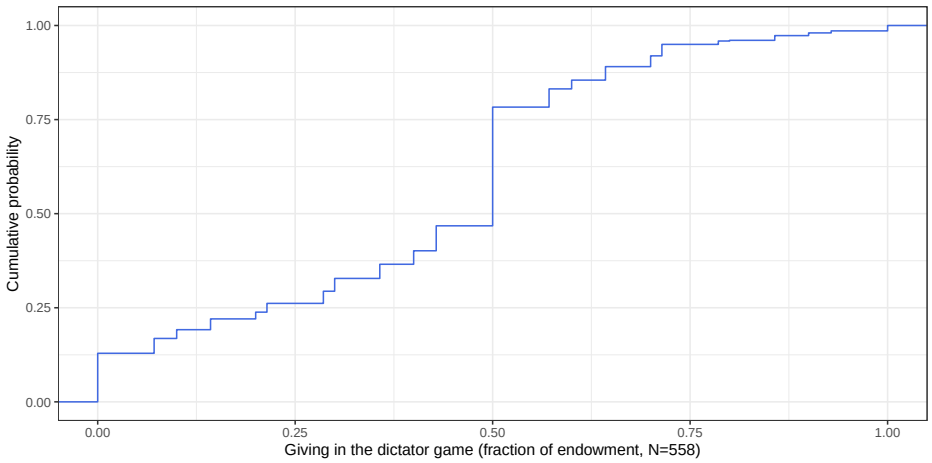
10. Although it is easy to distinguish Italian and non-Italian sounding names, to make sure not to misclassify, I asked the laboratory manager who was native Italian to check participants' first names after each session.

Table 1: Dictators' and paired recipients' characteristics, proximity between dictators and paired recipients, and dictator game allocation

	Female dictators		Male dictators		Difference (Female – Male)		
	Mean	SD	Mean	SD	Mean	SE	P-value
Panel A: Own characteristics							
IQ level	6.52	1.20	6.89	1.24	-0.37	0.18	0.04
IQ rank	3.83	1.59	3.31	1.73	0.52	0.24	0.03
Age	23.68	2.62	23.23	2.81	0.45	0.39	0.25
From Emilia-Romagna	0.18	0.39	0.19	0.39	0.00	0.06	0.94
Humanities	0.58	0.50	0.32	0.47	0.26	0.07	0.00
Social sciences	0.15	0.36	0.24	0.43	-0.09	0.06	0.13
STEM	0.27	0.45	0.44	0.50	-0.17	0.07	0.01
Post bachelor	0.53	0.50	0.37	0.49	0.16	0.07	0.03
Overconfidence	0.31	0.78	0.56	0.72	-0.25	0.11	0.02
Time on feedback (sec.)	107.67	89.88	107.52	102.26	0.16	13.88	0.99
Observations	104		91				
Panel B: Paired recipients' characteristics							
IQ level	6.77	1.19	6.91	1.12	-0.14	0.09	0.11
IQ rank	3.39	1.75	3.45	1.74	-0.05	0.10	0.61
Higher IQ	0.57	0.50	0.48	0.50	0.09	0.05	0.08
Age	23.17	2.57	23.55	2.98	-0.37	0.24	0.12
Female	0.50	0.50	0.43	0.50	0.07	0.04	0.06
From Emilia-Romagna	0.15	0.36	0.25	0.43	-0.09	0.04	0.01
Observations	298		260				
Panel C: Social distance with paired recipients							
Did not know at all	0.98	0.15	0.95	0.23	0.03	0.02	0.14
Knew but not well	0.02	0.15	0.03	0.18	-0.01	0.02	0.48
Saw before	0.00	0.00	0.02	0.14	-0.02	0.01	0.06
Observations	298		260				
Panel D: Dictator game allocation (fraction of endowment)							
Allocation	0.43	0.22	0.37	0.25	0.06	0.03	0.04
Allocation (residualized)	0.03	0.22	-0.03	0.25	0.06	0.03	0.06
Observations	298		260				

Notes: This table shows dictators' (Panel A) and paired recipients' characteristics (Panel B), social distance between dictators and paired recipients (Panel C), and dictators' allocation (Panel D) separately for female and male dictators. Residualized allocation is residual from the regression of the dictator game allocation as a fraction of endowment on IQ fixed effects, and shows within dictator IQ variation. P-values for the difference between female and male dictators are calculated with heteroskedasticity-robust standard errors with Bell and McCaffrey (2002)'s small sample bias adjustment for Panel A and with Pustejovsky and Tipton (2018)'s small cluster bias adjustment for Panels B-D.

Figure 2: CDF of the dictators' allocation



Notes: These figures plot the empirical CDF of the dictators' allocation as a fraction of endowment and show that the CDF resembles that of Bohnet and Frey (1999)'s one-way identification with information treatment where the social distance between dictators and recipients is the closest to my setting.

4 The role of gender and IQ on dictators' allocation

In this section, I document evidence that one's gender, IQ, or their interaction do not affect the allocation they receive from dictators, both in mean and distribution. If anything, women with higher IQ receive more allocation. I also document evidence that participants' belief about paired recipients' IQ is roughly consistent with the experimental design.

4.1 The role of gender and IQ on dictators' allocation: Estimating equation

I estimate the following model with OLS:

$$Allocate_{ij} = \beta_1 HigherIQ_{ij} + \beta_2 Female_j + \beta_3 HigherIQ_{ij} \times Female_j + X'_{ij}\gamma + \mu_i^{IQ} + \epsilon_{ij} \quad (1)$$

where each variable is defined as follows:

- $Allocate_{ij} \in [0,1]$: dictator i's allocation to recipient j as a fraction of endowment.
- $HigherIQ_{ij} \in \{0,1\}$: an indicator variable equals 1 if recipient j's IQ is higher than that of dictator i.
- $Female_j \in \{0,1\}$: an indicator variable equals 1 if recipient j is female.
- X_{ij}: a set of additional covariates to increase statistical power and to address the potential ex-post imbalance. Online Appendix B provides a full description of the covariates.
- ϵ_{ij}: omitted factors that affect dictator i's allocation to recipient j conditional on covariates.

and $\mu_i^{IQ} \equiv \sum_{k=1}^{9} \mu^k \mathbb{1}[\text{i's IQ} = k]$ is fixed effects for the dictators' IQ (number of IQ questions they have solved correctly), where $\mathbb{1}$ is the indicator variable. Standard errors are clustered at the dictator level with Pustejovsky and Tipton (2018)'s small cluster bias adjustment.

Dictator's IQ fixed effects are included following Zimmermann (2020) so that the coefficients in

equation 1 capture allocation differences due to the recipients' IQ, not that of the dictators. Indeed, Online Appendix Table C1 shows that dictator IQ rank is uncorrelated with dictator characteristics conditional on dictator IQ fixed effects.

The key identification assumption is that conditional on dictator IQ fixed effects, recipient gender, recipient's IQ rank relative to dictator's, and their interaction are uncorrelated with factors that affect dictator game allocation. The recipient's gender is ex-ante exogenous to dictator game allocation by random desk assignment. Recipient's IQ rank is also ex-ante exogenous to dictator game allocation conditional on dictator's IQ fixed effects by random desk assignment and random matching of dictators and recipients in part 2. Online Appendix Table C2 shows that they are indeed uncorrelated with the dictator or the paired recipient characteristics, dictator game rounds, or social distance between dictators and paired recipients.

4.2 The role of gender and IQ on dictators' allocation: Results

Regression results Columns 1-5 of Table 2 present the regression results of equation 1. Columns 1 and 2 show that when we do not control for dictators' IQ, dictators allocate more to higher IQ recipients although the difference is statistically insignificant: lower IQ dictators allocate more to higher IQ recipients. Columns 2-5 gradually add more controls and show that coefficient estimates are stable across different specifications, suggesting irregularities in the data is unlikely to be driving the results.

Looking at column 5, the coefficient estimates on all covariates are statistically insignificant even at 10%. They are quantitatively insignificant as well: the effect size of typical dictator game experiments that examine the role of social distance with university students is 8.9% to 11.42% of the endowment,[11] which is much larger than the effect sizes in column 5 that ranges from 0.6% to 3.5% of the endowment. If anything, the coefficient estimate on the interaction between higher IQ recipient and female recipient may be quantitatively significant: female recipients who happen to have a higher IQ than dictators receives about 3.5 percentage point more than equivalent male recipients, albeit statistically insignificant. The same results hold when we separately examine female (column 6) and male (column 7) dictators.

Addressing "beauty premium" Note that the so-called beauty premium – that people are more generous to physically attractive people (Landry et al. 2006) and hence affects dictators' allocation (Rosenblat 2008) – does not confound the results even if it is gender-specific (e.g., women smile more on a photo and hence look more approachable). It is because I am comparing recipients of the same gender who happen to have a higher IQ than their dictators and happen to have a lower IQ than their dictators; thus, gender-specific beauty premium is kept constant.

11. For example. Charness and Gneezy (2008) examine how informing the recipient's family name increases the dictators' allocation using a university student sample and find an 8.9% increase in allocation as a fraction of endowment. Leider et al. (2010) find using a university student sample that dictators increase allocation by 11.42% as a fraction of endowment for their friends relative to someone living in the same student dormitory. Brañas-Garza et al. (2010) also find using a university student sample that dictators give about 10% more of their endowment to friends relative to other students in the same class.

Table 2: The role of gender and IQ in dictators' allocation

Outcome:	Dictator's allocation (fraction of endowment)			
Sample:	All dictators			
	(1)	(2)	(3)	(4)
Higher IQ recipient	0.031	0.011	0.013	0.005
	(0.031)	(0.033)	(0.033)	(0.033)
	[-0.030, 0.093]	[-0.054, 0.075]	[-0.053, 0.078]	[-0.059, 0.070]
Female recipient	0.018	0.014	0.014	0.007
	(0.027)	(0.027)	(0.027)	(0.026)
	[-0.037, 0.072]	[-0.040, 0.067]	[-0.040, 0.068]	[-0.044, 0.058]
Higher IQ recipient x Female recipient	0.024	0.027	0.026	0.034
	(0.037)	(0.037)	(0.037)	(0.036)
	[-0.048, 0.097]	[-0.045, 0.100]	[-0.048, 0.099]	[-0.037, 0.105]
Dictator IQ FE	-	✓	✓	✓
Round FE	-	-	✓	✓
Social distance FE	-	-	✓	✓
Dictator controls	-	-	-	✓
Recipient controls	-	-	-	-
Higher IQ recipient x Female recipient +Female recipient	0.042	0.041	0.040	0.041
	(0.026)	(0.026)	(0.026)	(0.026)
	[-0.009, 0.093]	[-0.010, 0.092]	[-0.012, 0.091]	[-0.010, 0.092]
Baseline Mean	0.373	0.373	0.373	0.373
Baseline SD	0.261	0.261	0.261	0.261
Adj. R-squared	0.006	0.010	0.006	0.047
Observations	558	558	558	558
Clusters	195	195	195	195

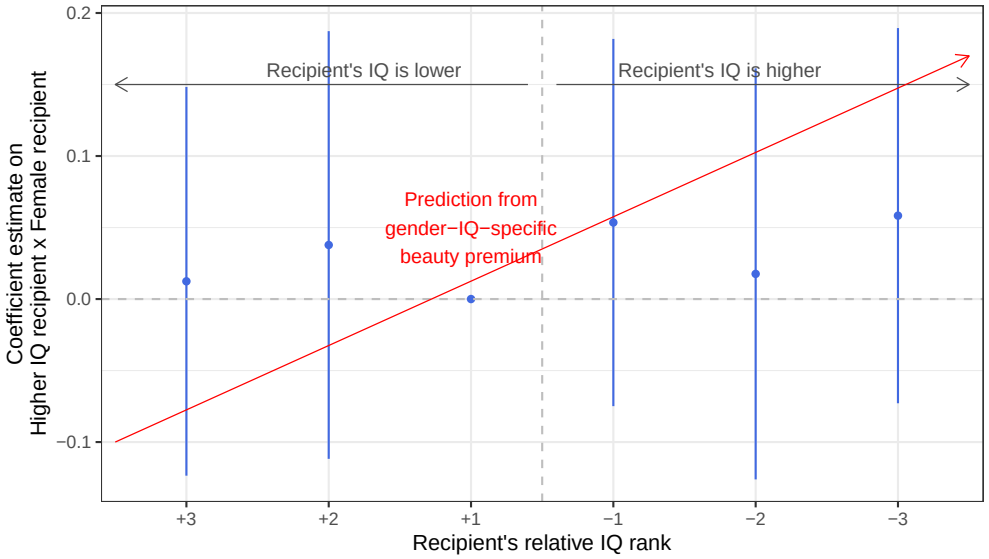
Outcome:	Dictator's allocation (fraction of endowment)			Belief on IQ (fraction of baseline SD)
Sample:	All dictators	Female dictators	Male dictators	All recipients
	(5)	(6)	(7)	(8)
Higher IQ recipient	0.006	-0.049	0.048	0.127
	(0.034)	(0.042)	(0.055)	(0.166)
	[-0.061, 0.072]	[-0.132, 0.034]	[-0.062, 0.158]	[-0.203, 0.458]
Female recipient	0.006	-0.014	0.014	-0.193
	(0.026)	(0.037)	(0.034)	(0.160)
	[-0.045, 0.057]	[-0.089, 0.061]	[-0.054, 0.082]	[-0.511, 0.124]
Higher IQ recipient x Female recipient	0.035	0.057	0.031	0.281
	(0.037)	(0.046)	(0.061)	(0.215)
	[-0.037, 0.107]	[-0.035, 0.148]	[-0.090, 0.152]	[-0.143, 0.706]
Dictator IQ FE	✓	✓	✓	✓
Round FE	✓	✓	✓	✓
Social distance FE	✓	✓	✓	✓
Dictator controls	✓	✓	✓	✓
Recipient controls	✓	✓	✓	✓
Higher IQ recipient x Female recipient +Female recipient	0.041	0.042	0.045	0.088
	(0.026)	(0.029)	(0.047)	(0.141)
	[-0.011, 0.093]	[-0.015, 0.100]	[-0.048, 0.138]	[-0.190, 0.365]
Baseline Mean	0.373	0.359	0.355	3.559
Baseline SD	0.261	0.256	0.262	1.000
Adj. R-squared	0.050	0.021	0.080	0.048
Observations	558	298	260	368
Clusters	195	104	91	193

Notes: This table presents the regression results of equation 1. Column 1 shows that when we do not control for dictators' IQ, dictators allocate more to higher IQ recipients. Columns 2-5 gradually add more controls and show that coefficient estimates are stable across different specifications, suggesting irregularities in the data is unlikely to be driving the results. Column 5 shows that the coefficient estimates on all covariates are statistically and quantitatively insignificant; if anything, the coefficient estimate on the interaction between the higher IQ recipient and the female recipient may be quantitatively significant. Columns 6 and 7 show that the same results hold when we separately examine female and male dictators. Column 8 shows beliefs about paired recipients' IQ is roughly consistent with the experimental design. The standard error (in parenthesis) and the 95% confidence interval (in bracket) are reported below each coefficient estimate. The standard errors are clustered at the dictator level with Pustejovsky and Tipton (2018)'s small cluster bias adjustment. Baseline mean and standard deviation are that of lower IQ male recipients. Significance levels: * 10%, ** 5%, and *** 1%.

One may even wonder whether higher IQ people are more physically attractive because they tend to earn more (Hamermesh and Biddle 1994) and look more confident (Mobius and Rosenblat 2006). However, if so, it is the premium they also receive in the real world, and controlling for that premium biases the results.

Yet, one may still wonder if beauty would appear more salient in the experimental setting than in the real world. To address this concern, Figure 3 plots the coefficient estimate on the interaction between higher IQ recipient and female recipient, separately for each IQ rank difference between dictator and recipient, along with the 95% confidence intervals. Note that the probability of receiving a better IQ rank (smaller IQ rank) is higher the higher one's IQ is. Thus, if IQ and beauty are positively correlated *and* more beauty recipients receive a higher allocation, then the coefficient estimate should be more positive/less negative the larger the IQ rank differences when the recipient's IQ is higher than that of dictator's, and should be less positive/more negative the larger the IQ rank differences when the recipient's IQ is lower than that of dictator's. Thus, in the presence of gender-IQ-specific beauty premium, we should expect an upward-sloping relationship between IQ rank and dictator game allocation, as shown in the red arrow. However, the estimates are inconsistent with the beauty premium prediction.

Figure 3: The role of gender and IQ in dictators' allocation – Addressing "beauty premium"

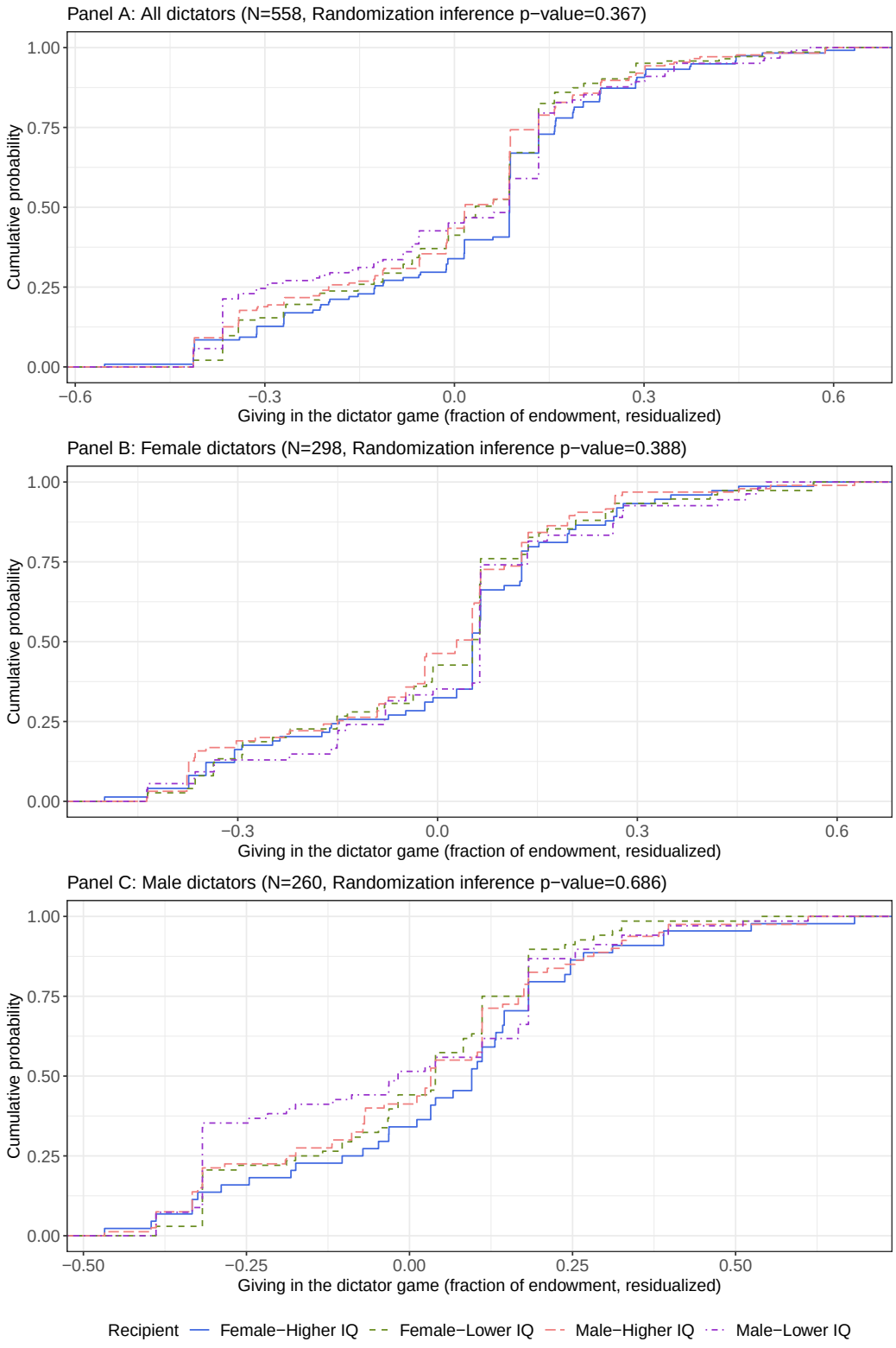


Notes: This figure plots the coefficient estimate on the interaction between higher IQ recipient and female recipient, separately for each IQ rank difference between dictator and recipient, along with the 95% confidence intervals. The red arrow is a relationship one should expect in the presence of gender-IQ-specific beauty premium. The standard errors are clustered at the dictator level with Pustejovsky and Tipton (2018)'s small cluster bias adjustment.

Distribution results While OLS only picks up the average effect, these results also hold in distribution. Panel A of Figure 4 presents empirical CDFs of dictators' allocation for each recipient type, residualized with the dictator IQ fixed effects to give a causal interpretation to the differences.[12]

12. Residualized allocation is residual from regression of dictators' allocation on dictator IQ fixed effects.

Figure 4: The role of gender and IQ in dictators' allocation – Distribution



Notes: These figures show the empirical CDF of residualized dictators' allocation by recipient types for all dictators (panel A), female dictators (Panel B), and male dictators (Panel C). The figures show the CDFs of dictators' allocation for each recipient type almost coincide and they are statistically indistinguishable from each other, even when we separately examine female and male dictators. The randomization inference p-value is calculated with the Kruskal-Wallis test.

The figure shows that the CDFs of dictators' allocation for each recipient type almost coincide. The randomization inference (Young 2019) using the Kruskal-Wallis test shows that the p-value of the differences in the CDFs is 0.37, which is far above the conventional 5% cutoff.[13] If anything, the CDF of higher IQ female recipients (the blue line) slightly lies on the right of the other CDFs across the x-axis values, suggesting they might receive a slightly higher allocation. The same results hold when we separately examine female (Panel B) and male (Panel C) dictators. Thus, one's gender, IQ, or their interaction do not affect the allocation they receive from dictators, both in mean and distribution.

Belief results To complement the findings so far, column 8 of Table 2 presents the regression results of equation 1 but with recipients' beliefs about paired recipients' IQ as the dependent variable. As discussed in section 2.1, random desk assignment ensures that recipients' belief proxies dictators' belief. Online Appendix Table C3 shows the ex-post balance of this comparability.

Column 8 shows that none of the coefficient estimates are statistically significant, may be because participants did not want to admit that their IQ is lower than the paired recipients even at the cost of reducing their payoff. However, the coefficient estimate on the higher IQ recipient is positive. Also, the coefficient estimate on the sum of the coefficient estimate on the female recipient and the interaction between the female recipient and the higher IQ recipient is positive. These suggest that participants correctly believe that male and female recipients with higher IQ solved a larger number of IQ questions. The coefficient estimate on female recipient is negative, suggesting that participants believe lower IQ female recipients solved a fewer IQ questions than lower IQ male recipients. Thus, participants' belief about paired recipients' IQ is roughly consistent with the experimental design.

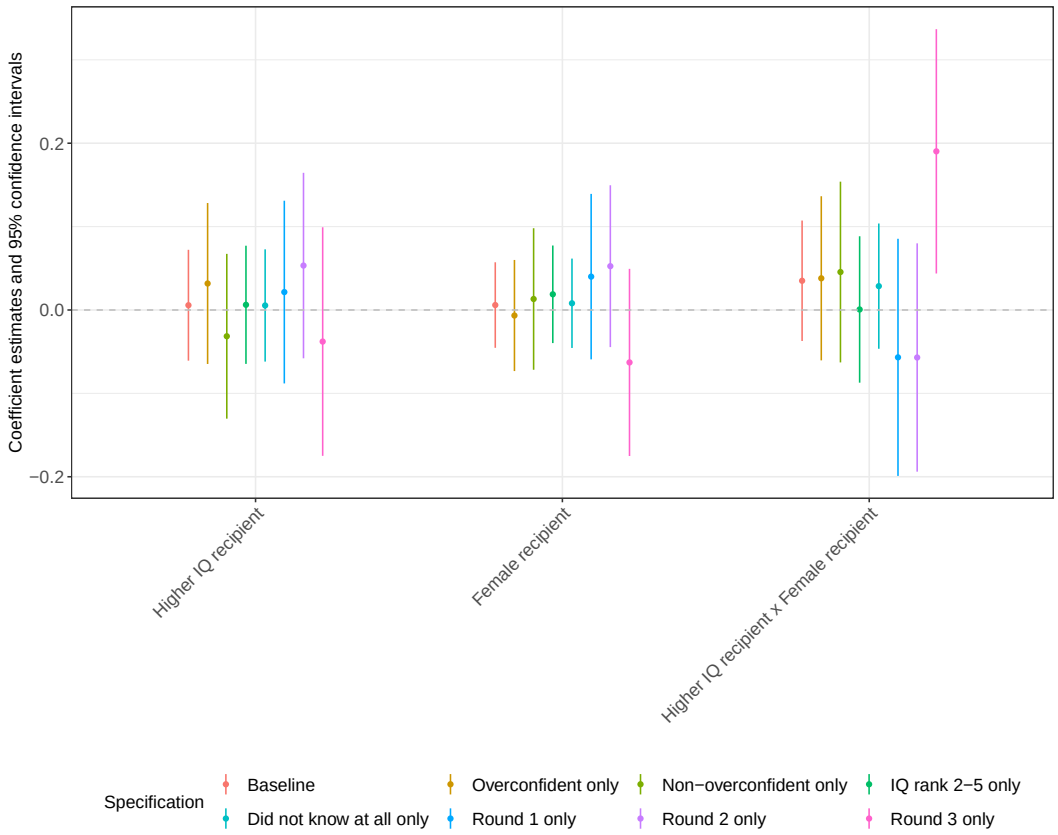
5 Robustness of the findings

In Figure 5, I re-estimate equation 1 with various sub-samples and plot the coefficient estimates along with their 95% confidence intervals to show the robustness of the findings in Table 2. I plot the estimate of column 5 of Table 2 with the red dot and line labeled as "Baseline" as a reference.

First, overconfident dictators may dislike higher IQ recipients more and hence allocate less. However, the estimates for overconfident (the brown dot and line) and non-overconfident dictators (the dark green dot and line) are very similar to the baseline estimates. Second, since dictators with IQ rank 1 only face lower IQ recipients and IQ rank 6 only face higher IQ recipients, they may behave differently from other dictators. However, the estimates with dictators of IQ rank 2-5 only (the green dot and line) provide very similar estimates as the baseline estimates. Third, although I excluded allocations where dictators knew the paired recipients "very well," knowing the paired recipients even a little may still affect the allocation. However, the estimates with allocations where dictators did not know at all the paired recipients (the light green dot and line) are very similar to the baseline estimates.

13. I use randomization inference to address arbitrary dependency among allocations. The null hypothesis is that all CDFs coincide.

Figure 5: The role of gender and IQ in dictators' allocation: Robustness



Notes: This figure re-estimates equation 1 with various sub-samples and plots the coefficient estimates along with their 95% confidence intervals to show the robustness of the findings in Table 2. The standard errors are clustered at the dictator level with Pustejovsky and Tipton (2018)'s small cluster bias adjustment for specifications from "Baseline" to "Did not know at all only" and heteroskedasticity-robust with Bell and McCaffrey (2002)'s small sample bias adjustment for specifications "round 1 only," "round 2 only," and "round 3 only."

Last, dictators play three-rounds of dictator games, and there can be across-round heterogeneity. The blue dot and line are estimates with round 1 only, the purple dot and line with round 2 only, and the pink dot and lines with round 3 only. There is indeed some heterogeneity; especially, in round 3, female recipients who happen to have a higher IQ than dictators receive statistically significantly higher allocation: they receive nearly 20 percentage point more allocation (as a fraction of endowment). It is unclear why dictators allocate higher in round 3; however, the bottom line is that it is consistent with that women with higher IQ receive more allocation, if anything. Also, it could be due to chance as I run several robustness regressions: for example, Gelman and Carlin (2014)'s Type M error ratio is about 2.5, suggesting that the estimate is likely to be 2.5 times larger than the true size.[14] Dividing the round 3 estimate by 2.5 makes it very close to the baseline estimate.

14. I use as the true value -0.47/(7+5+7)*3≈-0.074 from the pre-analysis plan (I divided -0.47 by the average of the dictator endowment).

6 Conclusion

This paper shows that gender, cognitive skills, or their interactions may not play a significant role in other people's generosity. If anything, people are more generous to women with high cognitive skills. While several studies show people perceive and treat women in traditionally male domains negatively (e.g., leadership, competition), and these domains typically require cognitive skills. My results suggest that these studies' findings are unlikely to be driven by violation of cognitive skill-related gender norms, which has implications for the role of gender norms in labor markets.

References

Aguiar, Fernando, Pablo Brañas-Garza, Ramón Cobo-Reyes, Natalia Jimenez, and Luis M. Miller. 2009. "Are Women Expected to Be More Generous?" *Experimental Economics* 12 (1): 93–98.

Aksoy, Billur, Ian Chadd, and Boon Han Koh. 2021. *(Anticipated) Discrimination against Sexual Minorities in Prosocial Domains.* Working Paper.

Almås, Ingvild, Alexander W. Cappelen, Kjell G. Salvanes, Erik Ø Sørensen, and Bertil Tungodden. 2016. "What Explains the Gender Gap in College Track Dropout? Experimental and Administrative Evidence." *American Economic Review: Papers & Proceedings* 106 (5): 296–302.

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

Autor, David, and Melanie Wasserman. 2013. *Wayward Sons: The Emerging Gender Gap in Labor Markets and Education.* Report. Washington, DC: Third Way.

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.

Bell, Robert M., and Daniel F. McCaffrey. 2002. "Bias Reduction in Standard Errors for Linear Regression with Multi-Stage Samples." *Survey Methodology* 28 (2): 169–181.

Bertrand, Marianne. 2011. "New Perspectives on Gender." In *Handbook of Labor Economics,* edited by David Card and Orley Ashenfelter, 4:1543–1590. Amsterdam, Netherlands: Elsevier.

———. 2018. "Coase Lecture – The Glass Ceiling." *Economica* 85 (338): 205–231.

Bilén, David, Anna Dreber, and Magnus Johannesson. 2021. "Are Women More Generous than Men? A Meta-Analysis." *Journal of the Economic Science Association.*

Bilker, Warren B., John A. Hansen, Colleen M. Brensinger, Jan Richard, Raquel E. Gur, and Ruben C. Gur. 2012. "Development of Abbreviated Nine-Item Forms of the Ravens Standard Progressive Matrices Test." *Assessment* 19 (3): 354–369.

Bohnet, Iris, and Bruno S. Frey. 1999. "Social Distance and Other-Regarding Behavior in Dictator Games: Comment." *American Economic Review* 89 (1): 335–339.

Bolton, Gary E., and Elena Katok. 1995. "An Experimental Test for Gender Differences in Beneficent Behavior." *Economics Letters* 48 (3): 287–292.

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.*

Boschini, Anne, Astri Muren, and Mats Persson. 2012. "Constructing Gender Differences in the Economics Lab." *Journal of Economic Behavior & Organization* 84 (3): 741–752.

Brañas-Garza, Pablo, Ramón Cobo-Reyes, María Paz Espinosa, Natalia Jiménez, Jaromír Kováík, and Giovanni Ponti. 2010. "Altruism and Social Integration." *Games and Economic Behavior* 69 (2): 249–257.

Bursztyn, Leonardo, Thomas Fujiwara, and Amanda Pallais. 2017. "'Acting Wife': Marriage Market Incentives and Labor Market Investments." *American Economic Review* 107 (11): 3288–3319.

Cappelen, Alexander, Ranveig Falch, and Bertil Tungodden. 2019. *The Boy Crisis: Experimental Evidence on the Acceptance of Males Falling Behind.* HCEO Working Paper 2019-014.

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.

Cawley, John, James Heckman, and Edward Vytlacil. 2001. "Three Observations on Wages and Measured Cognitive Ability." *Labour Economics* 8 (4): 419–442.

Ceci, Stephen J., and Wendy M. Williams. 2015. "Women Have Substantial Advantage in STEM Faculty Hiring, except When Competing against More-Accomplished Men." *Frontiers in Psychology* 6:1532.

Chakraborty, Priyanka, and Danila Serra. 2021. *Gender and Leadership in Organizations: Promotions, Demotions and Angry Workers.* Working Paper.

Charness, Gary, and Uri Gneezy. 2008. "What's in a Name? Anonymity and Social Distance in Dictator and Ultimatum Games." *Journal of Economic Behavior & Organization* 68 (1): 29–35.

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.

Chowdhury, Subhasish, Philip Grossman, and Joo Young Jeon. 2019. *Gender Differences in Giving and the Anticipation-about-giving in Dictator Games.* Economics & Management Discussion Paper 2019-13. Henley Business School, Reading University.

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.

Cunha, Flavio, and James J. Heckman. 2008. "Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Journal of Human Resources* 43 (4): 738–782.

Datta Gupta, Nabanita, Anders Poulsen, and Marie Claire Villeval. 2013. "Gender Matching and Competitiveness: Experimental Evidence." *Economic Inquiry* 51 (1): 816–835.

Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G. Wagner. 2011. "Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences." *Journal of the European Economic Association* 9 (3): 522–550.

Dreber, Anna, Tore Ellingsen, Magnus Johannesson, and David G. Rand. 2013. "Do People Care about Social Context? Framing Effects in Dictator Games." *Experimental Economics* 16 (3): 349–371.

Eagly, Alice H., and Steven J. Karau. 2002. "Role Congruity Theory of Prejudice toward Female Leaders." *Psychological Review* 109 (3): 573–598.

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

Eil, David, and Justin M. Rao. 2011. "The Good News-Bad News Effect: Asymmetric Processing of Objective Information about Yourself." *American Economic Journal: Microeconomics* 3 (2): 114–138.

Falk, Armin, Fabian Kosse, Pia Pinger, Hannah Schildberg-Hörisch, and Thomas Deckers. 2021. "Socioeconomic Status and Inequalities in Childrens IQ and Economic Preferences." *Journal of Political Economy* 129 (9): 2504–2545.

Gelman, Andrew, and John Carlin. 2014. "Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors." *Perspectives on Psychological Science* 9 (6): 641–651.

Goldin, Claudia. 2014. "A Grand Gender Convergence: Its Last Chapter." *American Economic Review* 104 (4): 1091–1119.

Gorsuch, Marina Mileo. 2019. "Gender, Sexual Orientation, and Behavioral Norms in the Labor Market." *Industrial and Labor Relations Review.*

Greiner, Ben. 2015. "Subject Pool Recruitment Procedures: Organizing Experiments with ORSEE." *Journal of the Economic Science Association* 1 (1): 114–125.

Håkansson, Sandra. 2021. "Do Women Pay a Higher Price for Power? Gender Bias in Political Violence in Sweden." *The Journal of Politics* 83 (2): 515–531.

Hamermesh, Daniel S., and Jeff E. Biddle. 1994. "Beauty and the Labor Market." *American Economic Review* 84 (5): 1174–1194.

Heckman, James J., Jora Stixrud, and Sergio Urzua. 2006. "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior." *Journal of Labor Economics* 24 (3): 411–482.

Heilman, Madeline E. 2001. "Description and Prescription: How Gender Stereotypes Prevent Women's Ascent Up the Organizational Ladder." *Journal of Social Issues* 57 (4): 657–674.

Heilman, Madeline E., Caryn J. Block, and Richard F. Martell. 1995. "Sex stereotypes: Do they influence perceptions of managers?" *Journal of Social Behavior and Personality* 10 (6): 237–252.

Heilman, Madeline E., and Tyler G. Okimoto. 2007. "Why Are Women Penalized for Success at Male Tasks?: The Implied Communality Deficit." *The Journal of Applied Psychology* 92 (1): 81–92. pmid: 17227153.

Heilman, Madeline E., Aaron S. Wallen, Daniella Fuchs, and Melinda M. Tamkins. 2004. "Penalties for Success: Reactions to Women Who Succeed at Male Gender-Typed Tasks." *The Journal of Applied Psychology* 89 (3): 416–427. pmid: 15161402.

Herrnstein, Richard J., and Charles Murray. 1996. *The Bell Curve: Intelligence and Class Structure in American Life.* New York, NY: Free Press.

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

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

Klinowski, David. 2018. "Gender Differences in Giving in the Dictator Game: The Role of Reluctant Altruism." *Journal of the Economic Science Association* 4 (2): 110–122.

Landry, Craig E., Andreas Lange, John A. List, Michael K. Price, and Nicholas G. Rupp. 2006. "Toward an Understanding of the Economics of Charity: Evidence from a Field Experiment." *The Quarterly Journal of Economics* 121 (2): 747–782.

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.

Leider, Stephen, Tanya Rosenblat, Markus M. Möbius, and Quoc-Anh Do. 2010. "What Do We Expect from Our Friends?" *Journal of the European Economic Association* 8 (1): 120–138.

Mobius, Markus M., and Tanya S. Rosenblat. 2006. "Why Beauty Matters." *American Economic Review* 96 (1): 222–235.

Moss-Racusin, Corinne A., Julie E. Phelan, and Laurie A. Rudman. 2010. "When Men Break the Gender Rules: Status Incongruity and Backlash against Modest Men." *Psychology of Men & Masculinity* (US) 11 (2): 140–151.

Niederle, Muriel, and Lise Vesterlund. 2011. "Gender and Competition." *Annual Review of Economics* 3 (1): 601–630.

Phelan, Julie E., Corinne A. MossRacusin, and Laurie A. Rudman. 2008. "Competent yet Out in the Cold: Shifting Criteria for Hiring Reflect Backlash Toward Agentic Women." *Psychology of Women Quarterly* 32 (4): 406–413.

Pustejovsky, James E., and Elizabeth Tipton. 2018. "Small-Sample Methods for Cluster-Robust Variance Estimation and Hypothesis Testing in Fixed Effects Models." *Journal of Business & Economic Statistics* 36 (4): 672–683.

Quadlin, Natasha. 2018. "The Mark of a Womans Record: Gender and Academic Performance in Hiring." *American Sociological Review* 83 (2): 331–360.

Ridgeway, Cecilia L. 2001. "Gender, Status, and Leadership." *Journal of Social Issues* 57 (4): 637–655.

Rosenblat, Tanya S. 2008. "The Beauty Premium: Physical Attractiveness and Gender in Dictator Games." *Negotiation Journal* 24 (4): 465–481.

Rudman, Laurie A. 1998. "Self-Promotion as a Risk Factor for Women: The Costs and Benefits of Counterstereotypical Impression Management." *Journal of Personality and Social Psychology* 74 (3): 629–645.

Rudman, Laurie A., and Kimberly Fairchild. 2004. "Reactions to Counterstereotypic Behavior: The Role of Backlash in Cultural Stereotype Maintenance." *Journal of Personality and Social Psychology* 87 (2): 157–176. pmid: 15301625.

Rudman, Laurie A., and Stephen E. Kilianski. 2000. "Implicit and Explicit Attitudes Toward Female Authority." *Personality and Social Psychology Bulletin* 26 (11): 1315–1328.

Rudman, Laurie A., Corinne A. Moss-Racusin, Julie E. Phelan, and Sanne Nauts. 2012. "Status Incongruity and Backlash Effects: Defending the Gender Hierarchy Motivates Prejudice against Female Leaders." *Journal of Experimental Social Psychology* 48 (1): 165–179.

Rudman, Laurie A., and Julie E. Phelan. 2008. "Backlash Effects for Disconfirming Gender Stereotypes in Organizations." *Research in Organizational Behavior* 28:61–79.

Williams, Wendy M., and Stephen J. Ceci. 2015. "National Hiring Experiments Reveal 2:1 Faculty Preference for Women on STEM Tenure Track." *Proceedings of the National Academy of Sciences* 112 (17): 5360–5365. pmid: 25870272.

Wuchty, Stefan, Benjamin F. Jones, and Brian Uzzi. 2007. "The Increasing Dominance of Teams in Production of Knowledge." *Science* 316 (5827): 1036–1039. pmid: 17431139.

Young, Alwyn. 2019. "Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results." *The Quarterly Journal of Economics* 134 (2): 557–598.

Zimmermann, Florian. 2020. "The Dynamics of Motivated Beliefs." *American Economic Review* 110 (2): 337–361.

Appendix

Appendix A The main change to the pre-analysis plan

In the initial design, recipients finished all the tasks except the post-questionnaire and left the laboratory before dictators received their IQ rank so that dictators could play the dictator game without recipients in the same room. The allocation to the recipients was paid electronically as a "participation fee" for the online post-questionnaire, which was sent to recipients via email after the session was over. However, as I ran the 1st session with this initial design with 24 participants, dictators had to wait idly for about 20-30 minutes until recipients left the laboratory, and dictators seemed to have lost concentration during this waiting time: about half of the dictators could not answer the comprehension questions about their IQ rank. Thus, I changed the design and let recipients stay in the laboratory while dictators played the dictator game. I looked at the 1st session data before making this change. I exclude the 1st session data in the analysis, but results are robust to including the 1st session data. The oTree code and instructions used for the 1st session are available upon request.

Appendix B Description of covariates

X_{ij} in equation 1 includes the following variables:

Dictator characteristics

- $Age_i \in \mathbb{N}$: dictator i's age.
- $Female_i \in \{0, 1\}$: an indicator variable equals 1 if dictator i is female, 0 otherwise.
- $FromEmiliaRomagna_i \in \{0, 1\}$: an indicator variable equals 1 if dictator i is from the Emilia-Romagna region where the experiment was conducted, 0 otherwise.
- $SocialSciences_i \in \{0, 1\}$: an indicator variable equals 1 if dictator i majors in social sciences, 0 otherwise.
- $STEM_i \in \{0, 1\}$: an indicator variable equals 1 if dictator i majors in natural sciences/mathematics, engineering, or medicine; 0 otherwise.
- $PostBachelor_i \in \{0, 1\}$: an indicator variable equals 1 if dictator i is either a master or post-bachelor student, a student in the 4th year or beyond in a bachelor-master combined program (bachelor is a 3 year program in Italy), or PhD student, 0 otherwise.
- $OverConfidence_i \in \{-1, 0, 1\}$: degree of dictator i's overconfidence. It is equal to -1 if dictator i's guess about the number of IQ test questions they have solved correctly is lower than the actual number, 0 if equal to the actual number, and 1 if higher than the actual number.

Recipient characteristics

- $Age_j \in \mathbb{N}$: recipient j's age.

- $FromEmiliaRomagna_j \in \{0, 1\}$: an indicator variable equals 1 if recipient j is from the Emilia-Romagna region where the experiment was conducted, 0 otherwise.

Fixed effects

- $\sum_{k=2}^{3} \mathbb{1}[\text{round}_{ij} = k]$: fixed effects for dictator game or belief elicitation round. $\mathbb{1}$ is the indicator variable.
- $\sum_{k=2}^{3} \mathbb{1}[\text{social distance}_{ij} = k]$: fixed effects for social distance between dictator i and recipient j. social distance$_{ij} = 1$ means dictator i did not know recipient j at all, $= 2$ knew but not well, and $= 3$ saw before. $\mathbb{1}$ is the indicator variable.

Appendix C Additional tables

Table C1: Exogeneity of dictator IQ rank conditional on dictator IQ fixed effects

Outcome:	Age	Female	From Emilia-Romagna	Human-ities	Social sciences	STEM	Post bachelor	Over-confidence
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
IQ rank = 2	0.010	0.221*	0.074	-0.095	0.034	0.061	0.151	0.146
	(0.796)	(0.128)	(0.104)	(0.130)	(0.088)	(0.130)	(0.127)	(0.200)
IQ rank = 3	-0.300	0.139	-0.007	-0.101	0.183	-0.081	0.183	0.160
	(0.776)	(0.143)	(0.103)	(0.142)	(0.120)	(0.137)	(0.137)	(0.241)
IQ rank = 4	-0.536	0.094	0.138	-0.146	0.101	0.045	0.187	0.430*
	(0.894)	(0.148)	(0.116)	(0.148)	(0.123)	(0.148)	(0.145)	(0.258)
IQ rank = 5	0.534	0.092	0.062	-0.220	0.166	0.054	0.061	0.158
	(0.959)	(0.165)	(0.128)	(0.175)	(0.128)	(0.165)	(0.156)	(0.271)
IQ rank = 6	-0.040	0.070	0.021	-0.368*	0.442***	-0.074	0.013	0.346
	(1.093)	(0.191)	(0.147)	(0.201)	(0.162)	(0.173)	(0.191)	(0.306)
Dictator IQ FE	✓	✓	✓	✓	✓	✓	✓	✓
F statistic	0.571	0.634	0.704	0.697	1.910*	0.626	0.739	0.830
Adj. R-squared	-0.012	0.016	-0.013	-0.010	0.024	0.011	-0.026	-0.020
Observations	195	195	195	195	195	195	195	195

Notes: This table shows dictator IQ rank is uncorrelated with dictator characteristics conditional on dictator IQ fixed effects. The F statistic shows the joint significance of IQ rank = 2 to IQ rank = 6 dummies. Heteroskedasticity-robust standard errors with Bell and McCaffrey (2002)'s small sample bias adjustment are reported below each coefficient estimate. Significance levels: * 10%, ** 5%, and *** 1%

Table C2: Exogeneity of the main regression's covariates conditional on dictator IQ fixed effects

Outcome:	Age	Female	From Emilia-Romagna	Human-ities	Social sciences	STEM	Post bachelor	Over-confidence
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Higher IQ recipient	-0.429	0.001	0.105**	-0.065	0.106**	-0.041	-0.071	0.063
	(0.350)	(0.064)	(0.048)	(0.065)	(0.051)	(0.060)	(0.063)	(0.107)
Female recipient	-0.228	0.060	0.080*	-0.026	0.015	0.011	-0.043	0.040
	(0.336)	(0.059)	(0.048)	(0.057)	(0.046)	(0.057)	(0.060)	(0.090)
Higher IQ recipient x Female recipient	0.431	0.010	-0.148**	0.014	-0.063	0.049	0.069	-0.051
	(0.458)	(0.082)	(0.064)	(0.081)	(0.062)	(0.079)	(0.084)	(0.129)
Dictator IQ FE	✓	✓	✓	✓	✓	✓	✓	✓
F statistic	0.522	1.078	2.074	0.505	1.731	0.661	0.417	0.119
Adj. R-squared	0.015	0.039	0.020	0.011	0.014	0.036	-0.000	-0.007
Observations	558	558	558	558	558	558	558	558
Clusters	195	195	195	195	195	195	195	195

Outcome:	Age (recipient)	From Emilia-Romagna (recipient)	Dictator game round 1	Dictator game round 2	Dictator game round 3	Did not know at all	Saw before	Knew but not very well
	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Higher IQ recipient	-0.792**	0.188***	-0.084	-0.026	0.110*	-0.002	0.008	-0.006
	(0.374)	(0.050)	(0.065)	(0.064)	(0.061)	(0.026)	(0.022)	(0.018)
Female recipient	-0.284	0.025	-0.084	0.037	0.047	0.020	-0.011	-0.009
	(0.344)	(0.038)	(0.062)	(0.058)	(0.059)	(0.020)	(0.017)	(0.010)
Higher IQ recipient x Female recipient	0.626	-0.100	0.137	-0.084	-0.053	-0.020	0.005	0.014
	(0.462)	(0.062)	(0.084)	(0.079)	(0.084)	(0.026)	(0.025)	(0.020)
Dictator IQ FE	✓	✓	✓	✓	✓	✓	✓	✓
F statistic	1.537	5.510***	0.941	0.890	1.207	0.666	0.415	1.071
Adj. R-squared	-0.002	0.027	-0.008	-0.009	-0.008	0.033	0.000	0.061
Observations	558	558	558	558	558	558	558	558
Clusters	195	195	195	195	195	195	195	195

Notes: This table shows recipient gender, recipient's IQ rank relative to dictator's, and their interaction are uncorrelated with dictator or paired recipient characteristics, dictator game rounds, or social distance between dictators and paired recipients. The F statistic shows the joint significance of all covariates. Cluster-robust standard errors with Pustejovsky and Tipton (2018)'s small cluster bias adjustment are reported below each coefficient estimate. Significance levels: * 10%, ** 5%, and *** 1%.

Table C3: Balance between dictators and recipients

	Recipients		Dictators		Difference (Recipients – Dictators)		
	Mean	SD	Mean	SD	Mean	SE	P-value
Panel A: Own characteristics							
IQ level	6.84	1.14	6.69	1.23	0.15	0.12	0.21
IQ rank	3.40	1.74	3.58	1.67	-0.18	0.17	0.30
Age	23.34	2.78	23.47	2.72	-0.14	0.28	0.63
From Emilia-Romagna	0.20	0.40	0.18	0.39	0.01	0.04	0.76
Humanities	0.34	0.48	0.46	0.50	-0.11	0.05	0.02
Social sciences	0.27	0.44	0.19	0.40	0.07	0.04	0.08
STEM	0.39	0.49	0.35	0.48	0.04	0.05	0.42
Post bachelor	0.49	0.50	0.46	0.50	0.04	0.05	0.48
Overconfidence	0.49	0.75	0.43	0.76	0.06	0.08	0.42
Time on feedback (sec.)	93.26	83.96	107.60	95.60	-14.34	9.13	0.12
Observations	193		195				
Panel B: Paired recipients' characteristics							
IQ level	6.84	1.16	6.84	1.16	0.00	0.07	1.00
IQ rank	3.42	1.74	3.42	1.74	0.00	0.09	0.97
Higher IQ	0.50	0.50	0.53	0.50	-0.03	0.04	0.52
Age	23.35	2.80	23.35	2.77	0.00	0.19	0.99
Female	0.47	0.50	0.47	0.50	0.00	0.03	0.99
From Emilia-Romagna	0.19	0.40	0.20	0.40	0.00	0.03	0.88
Observations	368		558				
Panel C: Social distance with paired recipients							
Did not know at all	0.98	0.14	0.96	0.19	0.02	0.01	0.15
Knew but not well	0.02	0.14	0.03	0.17	-0.01	0.01	0.39
Saw before	0.00	0.00	0.01	0.09	-0.01	0.00	0.06
Observations	368		558				
Panel D: Belief on paired recipient's IQ level (fraction of baseline SD)							
Belief on IQ level	3.48	1.04					
Belief on IQ level (residualized)	0.00	1.02					
Observations	368						

Notes: This table shows that recipients and dictators are comparable also ex-post. P-values for the difference between recipients and dictators are calculated with heteroskedasticity-robust standard errors with Bell and McCaffrey (2002)'s small sample bias adjustment for Panel A and with Pustejovsky and Tipton (2018)'s small cluster bias adjustment for Panels B-D.

Appendix D Experimental instructions

To the experimenter:

- *Before subjects arrive:*
- *Clear image cache from the browser.*
- *Put on each desk (i) a scratch paper and (ii) a pencil.*
- *Have a printed instructions ready.*
- *Set up photo booth. The brightness of the camera should be 172 and resolution 0.7 mb with 4:3 aspect ratio.*
- *Leave a paper in which participants write down their desk number on the photo booth.*
- *After registration:*
- *Give them photo taking instructions.*
- *Ask them to take photo at the photo booth, then take seat.*
- *After subjects took photo:*
- *Check that all the participants' photos are neutral: they must not signal nothing other than their gender.*
- *Make sure that the photos are saved as Pxx.jpg where xx is participant's desk number.*
- *After reserve participants left the room:*
- *Rename the photo name to the new desk number's for those who moved to new desks.*
- *Store photos in _static/photo folder.*
- *Startup Chrome & oTree*

App: personal_info

Page: DeskNumber

Please enter your desk number and click "Next"

[Your desk number:]

Page: PersonalInfo

Please check that the photo is yours

[Participant's photo]

The photo you took is displayed above. Please check that the photo is yours. Please also enter your first name. We will come to each desk and check the photo and the first name.

[Your first name:]

[Digital signature (please wait for us to sign you in):]

To the experimenter: before type in the password, do the followings:

- *Check that the photo and the first name correspond to the participant.*

Then click "Next" to let participants to proceed.

Page: Introduction

To the experimenter: read the instructions aloud.

Welcome!

You are participating in a study of the BLESS. For your participation, you will receive a fixed amount of [Participation fee]€. There are 2 parts in which you can earn additional earnings. The expected length is 1 hour.

During the study, we use your photo and first name to identify you. Your photo and the first name will only be used in this session and deleted immediately afterwards. However, we may ask some of you to allow us to use their photo in another study, which you can opt out.

The study is computerized, meaning that the computer program will give you precise instructions in each task. In the following you will find general instructions of the study, which you can always find in the bottom of the screen.

General instructions

- Please turn off your mobile phone.
- Please do not communicate with other participants.
- Please only use paper and pencil.
- Once you understand the instructions or enter your decisions, please click "Next" to proceed unless instructed otherwise.
- If you have any questions, please raise your hand at any time.

If there is no question, we will start the study.

To the experimenter:

- *Confirm that everyone turned off their mobile phone.*
- *Then, if there is no question, click "Advance slowest user(s)."*

After that, just sit in the experimenter area unless someone raises her or his hand. Do not read instructions aloud unless this document says to do so.

App: iqtest

Page: Introduction

Part 1: Instructions

In part 1, you will work on an IQ test, which is frequently used to measure intelligence. The IQ test you will work on is the Raven's Standardized Progressive Matrices Test.

You will solve the IQ test as follows: for each question, you will see an image in which a piece is missing. Below the image there will be several options. Choose the correct option among them to complete the image. There will be only one correct option.

An example is provided below. In the image, there are 9 large white squares each containing a small black square. In the first column, the small black square is located on the left; in the second column, in the middle; in the third column, on the right. In the first row, it is located on the top; in the second row, in the middle; in the third row, in the bottom. Thus, in the third

column of the third row, the small black square must be located in the right bottom, thus the correct option is 5.

[Raven matrix no. 31 here]

There are 9 questions in total and you have 9 minutes. Once the time is over, you will automatically be directed to the next page. You will earn [Payoff per IQ test]€ for each correct answer. There is no penalty for wrong answers. You can use paper and pencil on your desk.

Page: IQTest (9 minutes)

Please complete the image by choosing the correct option

[Raven IQ test]

Page: Guess

Guess the number of questions you solved

The IQ test is over.

We have randomly formed a group of 6 participants including you in this room and constructed a ranking among the 6 group members based on their IQ test performance.

A group member with rank 1 performed the best in the IQ test, followed by a group member with rank 2, 3, 4, 5, and 6. In case of a tie between group members, the computer randomly decided who receives the higher rank.

How many questions do you think you have solved correctly? If your guess is correct, you will additionally earn [Earnings from guess]€.

[Guess]

[Dictator] ***Page: Feedback***

Feedback

Among your 6 group members including you, you received **Rank [Participant's rank]**.

[Among your 6 group members, how many people performed better than you in the IQ test?:]

[Among your 6 group members, how many people performed worse than you in the IQ test?:]

App: dictator

[Dictator] ***Page: IntroductionDict***

Part 2: Instructions

In this part, half of you will be active participants who will work on the task described in the next page, and the remaining half will be passive participants who will NOT work on the task described in the next page.

[Dictator] ***Page: IntroductionDictCont***

Part 2: Instructions

You are assigned to a role of **active participant**.

Part 2 consist of 3 rounds. In each round, you will first receive an endowment (money). After that, you will be paired with a passive participant in your group.

Your task in this part is to allocate the endowment to yourself and the paired passive participant. The passive participants, other active participants, or anyone else other than us will never know who allocated how much.

At the end of the study, the computer will randomly select 1 out of 3 rounds and the amount you allocated to you in that round will be your earnings in this part.

The computer will also randomly select 1 out of 3 rounds for the paired passive participants and the amount you allocated to him or her in that round will be his or her earnings in this part.

[Recipient] ***Page: IntroductionRecip***

Part 2: Instructions

In part 2 consists of 2 rounds. In each round, you will be paired with another participant in your group.

Your task in this part is to guess how many questions the paired participant has solved correctly in the IQ test. For each correct guess, you will earn [Earning from guess other]€.

[Dictator] ***Page: PrepEndow***

Round [Round number] of 3

Please wait.

[Dictator] ***Page: OfferDict1-3***

Round [Round number] of 3

[Paired participant's photo]

[Paired participant's first name]

Rank [Paired participant's rank]

You have received **[7/5/7]€** for this round.

You have been paired with **[Paired participant's first name].**

Please allocate the endowment between yourself and [Paired participant's first name]. When you click the line below, a cursor appears. You can move the cursor by dragging it. Please move the cursor to your preferred position to determine the allocation.

[Slider from 0 to endowment that moves with increment of 0.5]

[Recipient] ***Page: GuessOther1-3***

Round [Round number] of 2

[Paired participant's photo]

[Paired participant's first name]

Rank [Paired participant's rank]

You have been paired with **[Paired participant's first name].**

How many questions do you think [Paired participant's first name] has solved correctly?

[Guess]

[Dictator] **Page: AnonymityCheckDict**

Round 3 of 3

Below we display the participants whom you were paired with. How well did you know him/her before participating in this study?

[Paired participant 1's photo]	[Paired participant 2's photo]	[Paired participant 3's photo]
[Paired participant 1's first name]	[Paired participant 2's first name]	[Paired participant 3's first name]
[I didn't know him/her at all, I saw him/her before, I knew him/her but not very well, I knew him/her very well]	[I didn't know him/her at all, I saw him/her before, I knew him/her but not very well, I knew him/her very well]	[I didn't know him/her at all, I saw him/her before, I knew him/her but not very well, I knew him/her very well]

[Recip] **Page: AnonymityCheckRecip**

Round 2 of 2

Below we display the participants whom you were paired with. How well did you know him/her before participating in this study?

[Paired participant 1's photo]	[Paired participant 2's photo]
[Paired participant 1's first name]	[Paired participant 2's first name]
[I didn't know him/her at all, I saw him/her before, I knew him/her but not very well, I knew him/her very well]	[I didn't know him/her at all, I saw him/her before, I knew him/her but not very well, I knew him/her very well]

Page: ShowResults

Results

The study is over. The results are provided below.

- In part 1, you solved [**Number of IQ test questions solved**] questions and earned [**Earnings from IQ test]€.** [If guess is correct] You have additionally earned **[Earnings from guess]€** because your guess about the number of questions solved was correct.
- [Dictator] In part 2, computer selected **round [1/2/3]** in which you allocated **[Allocation to self]€** to yourself.
- [Recipient] In part 2, you made **[Number of correct guesses on others]** guesses correct. So you earned **[Earnings from guesses other]€**.
- [Recipient] You additionally earned a top-up of **[Allocation from dictator]€**.

So, your total earnings are **[Participant's earnings]€** including [Participation fee]€ of participation fee.

Thank you for participating in this study! We will prepare your payment soon. Meanwhile, please answer a short questionnaire by clicking "Next." Your answer will be kept anonymous and will not affect your payment.

Page: Questionnaire1

Questionnaire 1 of 3

[Your study program: Agricultural and Food Sciences; Economics and Management; Education; Engineering and Architecture; Humanities; Languages and Literatures, Interpreting and Translation; Law; Medicine; Pharmacy and Biotechnology; Political Sciences; Psychology; Sciences; Sociology; Sport Sciences; Statistics; Veterinary Medicine]

[Please also type your full study program name in Italian:]

If you are enrolled in a specialized or professional program, please choose the closest study program. If you are enrolled in a post-bachelor vocational program, please choose the study program of your bachelor's degree. If you are an exchange student, please choose the study field closest to the one in your home university.

[Your degree program: Bachelor, Master/Post-bachelor, Bachelor-master combined (ciclo unico), Doctor]

[Your year in the degree program: 1st, 2nd, 3rd, 4th, 5th, 6th, 7th, 8th]

[Your age:]

[Your gender: Male, Female]

[Are you from Emilia-Romagna region?: Yes, No]

[Recipient] In another study, we'd like to use your photo. We will show your photo to some people in the University of Bologna only in this room, but no other people except us will see

your photo. Your photo will be deleted immediately after we finish another study. For your cooperation, we will pay you gratuity of [Gratuity for photo use]€. May we use your photo in another study?

[Yes, I allow the researcher to use my photo in another study; No, I do NOT allow the researcher to use my photo in another study]

[What do you think the study you participated was about?]

[Was there anything unclear or confusing about the study you participated?]

[Do you have any other comments? (optional)]

Page: Questionnaire2

To the experimenter:

- *Prepare payment.*

Questionnaire 2 of 3

Below we display the participants whom you were paired with. How well did you know him/her before participating in this study?

[Dictator]

[Paired participant 3's photo]	[Paired participant 1's photo]	[Paired participant 2's photo]
[Paired participant 3's first name]	[Paired participant 1's first name]	[Paired participant 2's first name]
[I didn't know him/her at all, I saw him/her before, I knew him/her but not very well, I knew him/her very well]	[I didn't know him/her at all, I saw him/her before, I knew him/her but not very well, I knew him/her very well]	[I didn't know him/her at all, I saw him/her before, I knew him/her but not very well, I knew him/her very well]

[Recipient]

[Paired participant 2's photo]	[Paired participant 1's photo]
[Paired participant 2's first name]	[Paired participant 1's first name]
[I didn't know him/her at all, I saw him/her before, I knew him/her but not very well, I knew him/her very well]	[I didn't know him/her at all, I saw him/her before, I knew him/her but not very well, I knew him/her very well]

Page: Questionnaire3

Questionnaire 3 of 3

[What do you think this study was about?]

[Was there anything unclear or confusing about this study?]

[Do you have any other comments? (optional)]

[Participants with payment less than 5€] ***Page: ExtraTask***

Extra task

Please solve the additions below and click next to earn [5€ – Participant's earnings]€.

84	33	64

[Sum of the above numbers:]

19	65	97

[Sum of the above numbers:]

Chapter 3

The Welfare Effects of Increased Legal Tolerance toward Domestic Violence

Yuki Takahashi*

Abstract

This paper studies how increased legal tolerance toward domestic violence affects married women's welfare using the domestic violence decriminalization bill introduced to the Russian national congress in 2016. Using difference-in-differences and flexibly controlling for macroeconomic shocks, I find that the bill decreased married women's life satisfaction and increased depression, especially among those with a college degree and a highly qualified white-collar occupation who are supposed to be more sensitive to gender regressive atmosphere. Consistent with this conjecture, people became more tolerant toward general and domestic violence after the bill. These findings suggest that the bill reduced married women's welfare partly through the gender regressive atmosphere.

JEL codes: J12, I31, K36, P37
Keywords: domestic violence, welfare, social norm, law, Russia

*Department of Economics, University of Bologna. Email: yuki.takahashi2@unibo.it. I am grateful to Anastasia Arabadzhyan, Sonia Bhalotra, Maria Bigoni, Margherita Fort, Angelina Nazarova, Olga Popova, and Vincenzo Scrutinio for helpful comments and discussions.

1 Introduction

Domestic violence leaves a long-lasting negative impact on women's (Delara 2016) and children's lives (Monnat and Chandler 2015). However, as we witnessed during the COVID-19 lockdown (Bhalotra et al. 2021a; Clerici and Tripodi 2021), domestic violence is quite prevalent across the world, both in developing and OECD countries (Devries et al. 2013; Garcia-Moreno et al. 2006). Despite the urgency to take action, several post-communist countries go against it: Poland is leaving the European treaty on violence against women[1] and Belarus is prosecuting female political activists,[2] to list a few. While international organizations express their concerns over such policies, there is not much empirical evidence on their consequences.

To fill the gap in the literature, this paper studies the effect of increased legal tolerance toward domestic violence on married women's welfare using the Russian domestic violence decriminalization bill. Russia introduced a bill to decriminalize some forms of domestic violence to the national congress in 2016, which was eventually enacted in 2017 (Isajanyan 2017). The bill has been criticized by several international organizations such as the United Nations[3] and the Human Rights Watch (n.d.), among others, as well as Russian NGOs and activists,[4,5,6] but it is still in force.

Using difference-in-differences with unmarried women as a control group and flexibly controlling for macroeconomic shocks, I find that the law decreased married women's life satisfaction and increased their depression level, with larger effects on women with a college degree and a highly qualified white-collar occupation who are presumably less prone to domestic violence but more sensitive to regressive gender unequal atmosphere. Consistent with this conjecture, people became more tolerant toward general as well as domestic violence after the bill. Taken together, the bill has likely reduced women's welfare partly because of the suppressive atmosphere it brought.[7]

This paper's main contribution is to the literature on the role of legal institutions on domestic violence: this paper provides evidence on how increased legal tolerance toward domestic violence affects married women's welfare in a country where women are highly educated and actively participate in labor force. While there is evidence that introducing a stricter law against domestic violence in a developing country decreases domestic violence incidence (Sanin 2021),[8] there is little evidence on the effects of a legal change in the opposite direction and in a country where women are actively participating in labor market. Aside from legal tolerance toward domestic violence, the studies find that women's higher legal power over marital relationships and reproductive issues reduces domestic violence. For example, Stevenson and Wolfers (2006) use the introduction of the

1. https://www.hrw.org/news/2020/07/28/poland-abandoning-commitment-women; the official name of the treaty is the "Council of Europe Convention on preventing and combating violence against women and domestic violence."
2. https://news.un.org/en/story/2021/10/1104092
3. https://www.themoscowtimes.com/2019/04/12/un-committee-sides-against-russia-in-first-domestic-violence-ruling-a65226
4. https://regnum.ru/news/society/2777954.html
5. https://time.com/5942127/russia-domestic-violence-women/
6. https://www.hrw.org/news/2017/01/23/russia-bill-decriminalize-domestic-violence
7. This also suggests that, since depression reduces one's economic outcomes (Ridley et al. 2020), the post-communist countries' regressive gender policies will likely affect their economy negatively.
8. Sanin (2021) uses criminalization of gender-based violence in Rwanda and the Rwandan Genocide.

US unilateral divorce law and find that the law decreased domestic violence. Corroborating this, Aizer and Dal Bó (2009) use the introduction of policies that prohibit women from withdrawing prosecution for their violent partner in California and find that the policies increased domestic violence reporting presumably because they worked as commitment devices for women's time-inconsistent preference.[9] However, another study finds contradictory results: Iyengar (2009) uses the introduction of a mandatory arrest law in some US states that required police to arrest reported abusers increased the probability that the abusers killed women due to a reduction in reporting and an increase in men's retaliation. On reproductive issues, Muratori (2021) uses abortion law change in Texas and finds that limiting access to abortion increases domestic violence and other violence against women.

This paper also relates to the literature on the effect of women's economic power on domestic violence. While studies also find that women's high economic power can backlash (Ericsson 2020; Erten and Keskin 2021; Tur-Prats 2019),[10] this strand of literature suggests that women's economic power reduces domestic violence. Leading evidence on this claim comes from Bhalotra et al. (2021b), who use Brazil's mass layoffs and examine women's and men's job loss separately. They find that women's job loss increases domestic violence, which suggests that women's economic power has stronger effect in reducing domestic violence than backlash effect. Corroborating this, Molina and Tanaka (2021) use increased demand for garment factory workers in Myanmar and find that women's increased paid employment opportunities reduce domestic violence. Outside the employment context, Haushofer et al. (2019) find that unconditional cash transfers to women reduced domestic violence in Kenya.

The remainder of the paper proceeds as follows. Section 2 provides details of the Russian domestic violence decriminalization bill. Section 3 describes data. Section 4 presents empirical strategy. Section 5 presents the results. Section 6 concludes.

2 Institutional context

Gender development in Russia and other post-communist countries Women in post-communist countries were once running far ahead of western counterparts for their labor force participation (e.g., see Boelmann, Raute, and Schönberg 2021). Even today, women in post-communist countries, and especially Russia, are highly educated and actively participating in labor market relative to men. Figure 1 plots the gender development index for Russia (blue), post-communist countries other than Russia (green), BRICS other than Russia (red), and OECD countries (purple) for the period from 1995 to 2019. Gender development index is a ratio of women's

9. That is, after their partner stops battering them, they will consider the cost of breaking up higher than the cost of receiving battery in the future.

10. Ericsson (2020) finds that an increase in women's potential earnings leads to higher domestic violence in Sweden and Erten and Keskin (2021) find that a decline in female employment reduces domestic violence using Syrian refugee arrivals in Turkey. Tur-Prats (2019) find that households in areas where mother-in-law takes care of some domestic work – hence women have more time to participate in the labor force – experience lower rate of domestic violence in Spain.

and men's human development index, which is calculated from mean and expected years of schooling, gross national income (GNI) per capita, and life expectancy at birth. Thus, the higher the index, the more women are educated, the more women earn, and the longer women live relative to men, with 1 being gender parity.

The figure shows that Russia had a very high gender development index at the beginning of 2000 – above gender parity and above OECD countries. Although their index has been gradually deteriorating since then, it is still higher than that of OECD and slightly above gender parity. Other post-communist countries also had a gender development index higher than that of OCED countries in 1995; although their index has not improved since then, their index is still very close to that of OECD. These are stark contrasts to the situation in other BRICS – a group of countries with a similar degree of economic development as Russia – whose gender development indexes have been very low, although they have been catching up OECD countries.[11]

Changes in battery penalties in the mid-2010s In July 2015, the Russian Supreme Court introduced a bill to make light battery[12] an administrative offense rather than a criminal offense to the Russian national congress (Isajanyan 2017). The bill initially included light battery against any person, including family members,[13] but before its implementation, the congress kept light battery against family members as a criminal offense.

However, the Russian Orthodox Church immediately made a statement against the exclusion of light battery toward family members from the bill, saying that it has "no moral justification and legal grounds" (Russian Orthodox Church 2016). Then a group of Russian national congress members introduced a bill to decriminalize light battery toward family members to the national congress in November 2016, which was enacted in February 2017. Figure 2 presents the timeline of the changes in light battery penalties and Table 1 presents changes in the penalties for various batteries. Table 2 presents the details of the penalties for various batteries shown in Table 1.

The politicians' and the Russian Orthodox Church's intention may have been to decriminalize battery against family members in general. However, what NGOs and activists were concerned about this bill was the decriminalization of domestic violence.[14,15] The bill has been criticized by several international organizations such as the United Nations[16] and the Human Rights Watch

11. Appendix Figure A1 presents women's (Panel A) and men's (Panel B) human development index for groups of countries included in Figure 1. The figure shows that women's human development index in Russia and other post-communist countries lags behind the OCED countries, likely because of the difference in the degree of economic development. However, compared to countries with a similar degree of economic development – other BRICS – women's human development in Russia and other post-communist countries is much higher.

12. Battery is defined as "Beatings or other violent actions that caused physical pain" (The Russian Federation 1996)

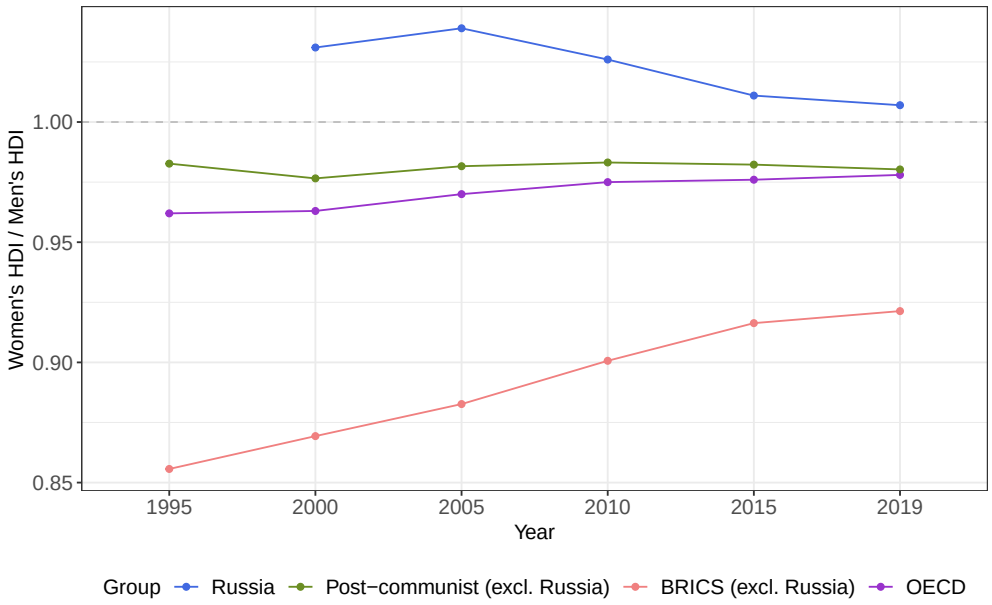
13. Family members are defined as "close relatives (husband, wife, parents, children, adoptive parents, adopted children, siblings, grandfathers, grandmothers, grandchildren), guardians , trustees, as well as persons who are in property with the person who committed the act provided for in this article, or persons who maintain a common household with him" (The Russian Federation 2016).

14. https://www.economist.com/europe/2017/01/28/why-russia-is-about-to-decriminalise-wife-beating

15. https://www.hrw.org/news/2017/01/23/russia-bill-decriminalize-domestic-violence

16. https://www.themoscowtimes.com/2019/04/12/un-committee-sides-against-russia-in-first-domestic-violence-ruling-a65226

Figure 1: Gender Development Index of Russia and groups of countries



Notes: This figure plots the gender development index for Russia (blue), post-communist countries other than Russia (green), BRICS other than Russia (red), and OECD countries (purple) for the period from 1995 to 2019. Gender development index is a ratio of women's and men's human development index, which is calculated from mean and expected years of schooling, gross national income (GNI) per capita, and life expectancy at birth. Thus, the higher the index, the more women are educated, the more women earn, and the longer women live relative to men, with 1 being gender parity. For the exact calculation of the index, see the technical notes of United Nations Development Programme (2020). Other BRICS are Brazil, India, China, and South Africa. Post-communist countries are initial Comecon members as defined in Britannica (*Comecon* 2019): former Soviet Union (Russia, Armenia, Azerbaijan, Belarus, Estonia, Georgia, Kazakhstan, Kyrgyzstan, Latvia, Lithuania, Moldova, Tajikistan, Turkmenistan, Ukraine, and Uzbekistan), Bulgaria, former Czechoslovakia (Czechia and Slovakia), Hungary, Poland, and Romania.
Source: UNDP Human Development Reports, Gender Development Index (http://hdr.undp.org/en/indicators/137906). Retrieved on December 26, 2021.

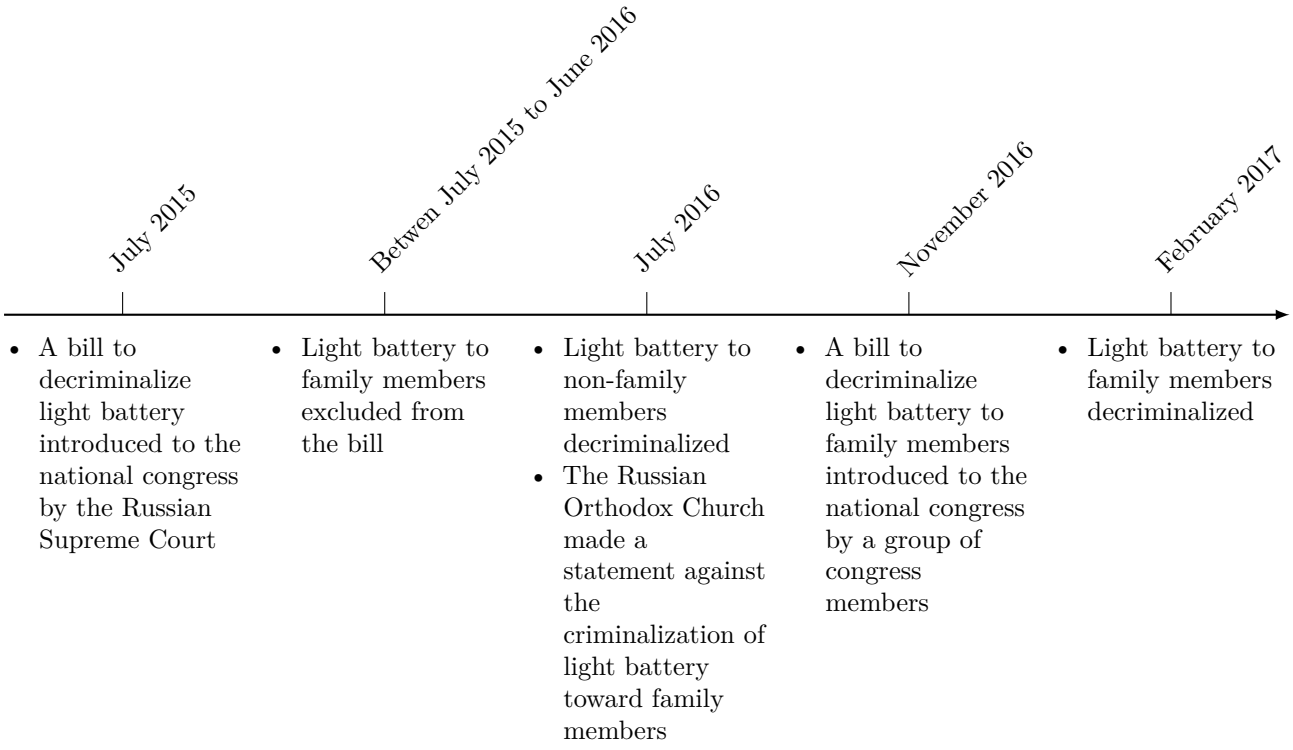
(n.d.), among others, as well as Russian NGOs and activists,[17,18,19] but it is still in force.

There are two things to note. First, these reforms are unlikely to be driven by that Russian people became more violent; rather, it is likely a part of a larger criminal law reform as the Supreme Court of the Russian Federation (2015) explains: it is an attempt for "humanization and liberalization of criminal legislation" of Russia. Figure 3 supports this claim: it plots the number of all registered crimes and serious crimes in Russia for the period from 2011 to 2019, normalized by their respective value in 2011. All (blue) includes all types of crimes, Murder (green) includes murders including attempts, Serious battery (red) includes batteries that result in serious injury, and Serious robbery (purple) includes stealing from someone with life-threatening means of violence.

The figure shows that registered cases for total crime and violent crimes have been decreasing from 2011 to 2015. While there is a jump in total crime in 2015, it is mainly driven by increase in theft that does not involve violence and fraud (Federal State Statistics Service 2017), neither of

17. https://regnum.ru/news/society/2777954.html
18. https://time.com/5942127/russia-domestic-violence-women/
19. https://www.hrw.org/news/2017/01/23/russia-bill-decriminalize-domestic-violence

Figure 2: Timeline of changes in light battery penalties in Russia



Notes: This figure shows the timeline of the changes in light battery penalties.
Sources: Isajanyan (2017), Human Rights Watch (n.d.), Russian Orthodox Church (2016), and The Russian Federation (2016, 2017).

Table 1: Changes in penalties for various batteries

	- July 2016	July 2016 - February 2017	February 2017 -
Battery to a family member 1st time in a given year	Criminal offense	Criminal offense (modified)	Administrative offense
Battery to a non-family member 1st time in a given year	Criminal offense	Administrative offense	
Battery to anyone 2nd time or more in a given year	Criminal offense	Criminal offense (modified)	
Battery to anyone that results in injury	Serious criminal offense		

Notes: This table shows changes in penalties for various batteries. Battery is defined as "Beatings or other violent actions that caused physical pain" (The Russian Federation 1996). Family member is defined as "close relatives (husband, wife, parents, children, adoptive parents, adopted children, siblings, grandfathers, grandmothers, grandchildren), guardians, trustees, as well as persons who are in property with the person who committed the act provided for in this article, or persons who maintain a common household with him" (The Russian Federation 2016).
Sources: Isajanyan (2017), Human Rights Watch (n.d.), and The Russian Federation (2016, 2017).

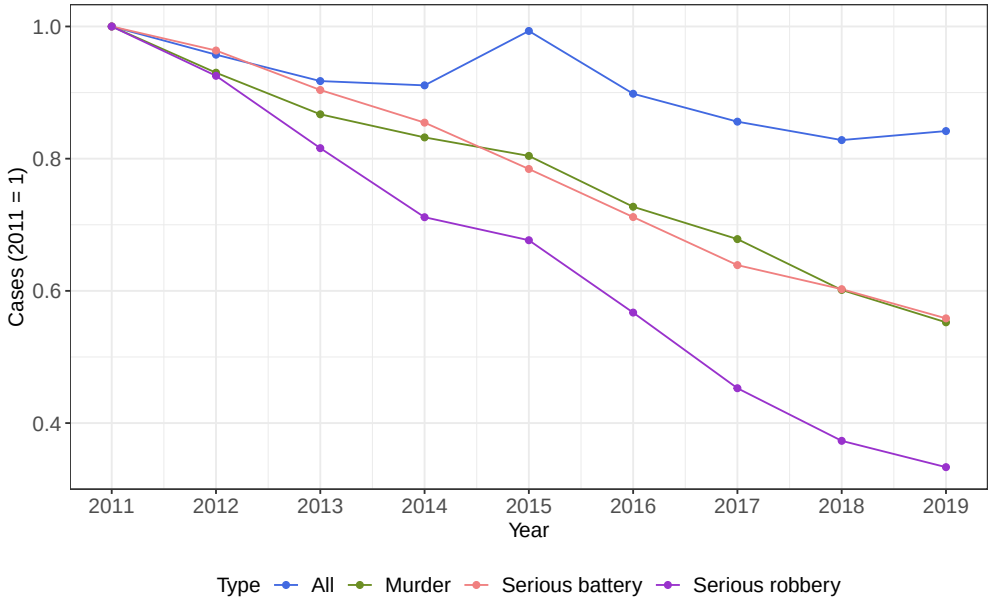
which is subject of the battery decriminalization bill and likely reflect a drop in the GDP growth in

Table 2: Details of the penalties for various batteries (one of the following applies)

	Administrative offense	Criminal offense	Criminal offense (modified)	Serious criminal offense
Fine (max.)	30000 rubles (≈400 USD)	40000 rubles (≈540 USD)		NA
Imprisonment (max.)	15 days	3 months		2 years
Labor (max.)	NA	6 months		2 years
Community service (max.)	120 hours	360 hours	240 hours	360 hours

Notes: This table presents the details of the penalties for various batteries shown in Table 1.
Sources: Isajanyan (2017), Human Rights Watch (n.d.), and The Russian Federation (2016, 2017).

Figure 3: Number of registered crimes in Russia by type (2011=1)



Notes: This figure plots the number of registered crimes by type in Russia for the period from 2011 to 2019, normalized by their respective value in 2011. All (blue) includes all types of crimes, Murder (green) includes murders including attempts, Serious battery (red) includes batteries that result in serious injury, and Serious robbery (purple) includes stealing from someone with life-threatening means of violence. Values in 2011 (in thousands): 2404.8 for All, 14.3 for Murder, 38.5 for Serious battery, and 20.1 for Serious robbery.
Sources: Federal State Statistics Service (2017, 2021).

2015.[20,21]

Second, the Russian legal stance against domestic violence was not necessarily looser than OECD countries as of the mid-2010s. For example, it was police officers' discretion whether a domestic

20. https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG?locations=RU
21. Although Federal State Statistics Service (2017) does not provide full breakdown of all crimes, the increase in all crimes from 2014 to 2015 is 197.9 thousands, the increase in theft that does not involve violence is 109.6 thousands, and the increase in fraud is 40.4 thousands.

violence abuser should be arrested in about half of the US states (American Bar Association 2014), and the European countries made a Europe-wide treaty against domestic violence only in 2011 (Council of Europe 2011). What makes this Russian case unique is that it sends an explicit message that some forms of domestic violence are not crimes. In what follows, I refer to the bill that decriminalized light battery toward family members as the "domestic violence decriminalization bill."

3 Data

Data on married women's welfare To examine the effect of the bill on married women's welfare, I use the Russia Longitudinal Monitoring Survey (RLMS), an individual and household panel survey data conducted every year by researchers at the Higher School of Economics of Moscow and the University of North Carolina at Chapel Hill (Kozyreva, Kosolapov, and Popkin 2016). The RLMS is a household-level nationally-representative annual survey where interviewers visit selected households and interview as many household members as possible. For household members of 13 years old and younger, the interviewers ask questions to the adult instead. From 2010 to 2013, above 6000 households and 16000 individuals were interviewed every year. The RLMS adds additional households each year to keep the number of households balanced.

The data contains information on individuals' health and welfare as well as demographics; I use it for the analysis for the period from 2011-2019 but exclude those who are added after 2015 because I could not define their marital status before the introduction of the domestic violence decriminalization bill.

Table 3 describes welfare measures that are my dependent variables (Panel A), demographic characteristics (Panel B), education level (Panel C), and occupation category (Panel D) for married (Treated) and unmarried women (Control) and their differences before the introduction of the domestic violence decriminalization bill (2011-2015). Marital status is defined as of 2015. Panel A shows average treated women have higher welfare than control women: they are more satisfied with their life and experience less depression than control women.[22,23]

Panel B shows that treated women are younger and more likely to be employed than control women. Treated women are also slightly more likely to be Russian Orthodox, although the difference is quantitatively small (3%). Panel C shows that treated women are more educated than control women. Corroborating this, Panel D shows that treated women are more likely to be in a higher paid occupation.

Thus, Panel A suggests that a simple comparison between treated and control women would not yield a causal effect of the bill and that the use of difference-in-differences would be appropriate.

22. English translation for the life satisfaction question is "satisfaction with life at present." The answer choices are "fully satisfied" being 1, "rather satisfied" being 2, "both yes and no" being 3, "less than satisfied" being 4, and "not at all satisfied" being 5. For ease of interpretation, I rescaled the answers into [0,1] interval and recoded it so that the higher the value, the more satisfied with the life.

23. English translation for the depression question is "had depression in last 12M?" and the answer choices are 1 being yes and 2 being no. I recoded this variable for ease of interpretation so that 0 being no and 1 being yes.

Table 3: Summary statistics for RLMS data: Treated vs. control women, 2011-2015

	Treated		Control		Difference (Treated – Control)		
	Mean	SD	Mean	SD	Mean	SE	P-value
Panel A: Welfare measures							
Life satisfaction (0-1)	0.60	0.25	0.51	0.28	0.09	0.01	0.00
Depression in the past 12 months (0/1)	0.11	0.31	0.15	0.35	-0.04	0.01	0.00
Panel B: Demographic characteristics							
Age	43.41	13.85	48.85	17.04	-5.45	0.39	0.00
Employed	0.64	0.48	0.51	0.50	0.13	0.01	0.00
Russian Orthodox	0.89	0.32	0.85	0.35	0.03	0.01	0.00
Panel C: Education							
Primary school or below	0.09	0.29	0.14	0.35	-0.05	0.01	0.00
Secondary school	0.57	0.49	0.60	0.49	-0.03	0.01	0.01
College or above	0.33	0.47	0.25	0.44	0.08	0.01	0.00
Panel D: Occupation category							
Managers/Professionals	0.58	0.49	0.52	0.50	0.06	0.01	0.00
Clerical/Services	0.28	0.45	0.31	0.46	-0.02	0.01	0.07
Blue-collar	0.09	0.29	0.14	0.34	-0.04	0.01	0.00
Agriculture/Craft	0.04	0.19	0.03	0.17	0.01	0.01	0.29
Military	0.00	0.07	0.00	0.03	0.00	0.00	0.00
Observations	15992		11815				
Individuals	3800		3066				

Notes: This table describes welfare measures that are my dependent variables (Panel A), demographic characteristics (Panel B), education level (Panel C), and occupation category (Panel D) for married (Treated) and unmarried women (Control) and their differences before the bill (2011-2015). Marital status is that of 2015. Occupation classification follows ISCO-08 (International Labour Office 2012) and defined as follows: Managers/Professionals (group 1, 2, and 3), Clerical/Services (group 4 and 5), Blue-collar (group 8 and 9), Agriculture/Craft (group 6 and 7), and Military (group 0). P-values of the difference between treated and control are calculated with standard errors clustered at the individual level.

However, Panels B-D suggests that macroeconomic shocks would likely have affected treated and control women differently and thus they are likely on a different time trend, invalidating standard difference-in-differences. My approach to deal with this possible differential time trend is to flexibly control for macroeconomic shocks at the region-education-occupation cell level.[24]

Data on people's attitudes To supplement the analysis of married women's welfare, I use the World Values Survey (WVS, Inglehart et al. 2020), a repeated cross-sectional nationally-representative survey that has been conducted since 1981 in more than 120 countries. The survey

24. Appendix Figure A2 presents simple year-by-year average and 95% confidence intervals of age at marriage (Panel A), employment status at marriage (Panel B), marriage rate (Panel C), and divorce rate (Panel D) around the bill, relative to the base year (2015). It shows that after the bill, women tend to postpone their marriage (Panel A). Although there is no significant change in employment status, marriage rate, or divorce rate (Panels B, C, D), the bill may have affected the composition of married women, thus I fix the marital status at the base year.

collects information on people's values in several dimensions such as "social, political, economic, religious and cultural values."[25] The main variables of interests are the answers to the question about (i) how justifiable it is "For a man to beat his wife" and (ii) how justifiable is"Violence against other people." I use the former as a proxy of an attitude toward domestic violence and the latter an attitude toward general violence. Around the domestic violence decriminalization bill, Russia was surveyed in 2011 and 2017. Thus, I use these two waves of the Russian survey. I include both women and men in the analysis.

Table 4: Summary statistics for WVS data: After vs. before the bill

	Post (2017)		Pre (2011)		Difference (Post – Pre)		
	Mean	SD	Mean	SD	Mean	SE	P-value
Panel A: Attitude measures							
Beating wife justifiable (0-1)	0.11	0.20	0.09	0.17	0.02	0.01	0.00
Violence against others justifiable (0-1)	0.11	0.19	0.07	0.15	0.04	0.01	0.00
Panel B: Demographic characteristics							
Age	43.17	15.17	44.04	15.77	-0.87	0.49	0.08
Female	0.58	0.49	0.54	0.50	0.03	0.02	0.04
Married	0.49	0.50	0.57	0.50	-0.08	0.02	0.00
Employed	0.64	0.48	0.64	0.48	0.00	0.02	0.78
Panel C: Education							
Primary school or below	0.10	0.30	0.01	0.07	0.09	0.01	0.00
Secondary school	0.23	0.42	0.67	0.47	-0.44	0.01	0.00
College or above	0.67	0.47	0.32	0.47	0.35	0.01	0.00
Observations	1699		2359				

Notes: This table describes attitude measures (Panel A), demographic characteristics (Panel B), and education level (Panel C) for everyone surveyed after (Post) and before the bill was introduced (Pre) and their differences. P-values of the difference between after and before are calculated with heteroskedasticity-robust standard errors.

Table 4 describes attitude measures (Panel A), demographic characteristics (Panel B), and education level (Panel C) for everyone surveyed after (Post) and before the domestic violence decriminalization bill was introduced (Pre) and their differences. Panel A shows people's attitude toward beating wife justifiable is very low in both periods, but has slightly increased after the bill.[26] Panel A also shows that people's attitude toward violence against others justifiable is very low too in both periods, but has increased after the bill.[27] However, the increase in attitude toward violence against others justifiable is slightly larger.

Panel B shows that there are 8% fewer married people and 3% more women in the post-period

25. https://www.worldvaluessurvey.org/WVSContents.jsp

26. The answer choices are 1-10 with 1 being "Never justifiable" and 10 being "Always justifiable." I rescaled the answers into [0,1] to make interpretation easier.

27. The answer choices are 1-10 with 1 being "Never justifiable" and 10 being "Always justifiable." I rescaled the answers into [0,1] to make interpretation easier.

survey, but age and employment status are not significantly different. Panel C shows that people are more educated in the post-period than in the pre-period.

4 Empirical strategy

I examine the effect of the domestic violence decriminalization bill on married women's welfare; I focus on married women because they are the group that is most exposed to domestic violence. I use unmarried women as a control group. Although unmarried women's welfare may have also been negatively affected by the bill through a drop in their expected utility from marriage, if anything, my estimate would be conservative.[28]

I consider the event year to be 2016 because (i) the Russian Orthodox Church already made a statement that the battery decriminalization should include domestic violence immediately after its enactment in July 2016 and it is very difficult to object to the Church,[29] (ii) the domestic violence decriminalization bill was introduced to the national congress in November 2016, and (iii) most 2016 data I use was collected from October to December 2016.[30]

Thus, I estimate the following event study form of the difference-in-differences equation via OLS using individual-level panel data with married women as a treated group and unmarried women as a control group, both defined as of 2015 to address the potential endogeneity of marital status to the bill:

$$Y_{it} = \sum_{l=2011, l\neq 2015}^{2019} \beta_l \mathbb{1}[t = l] \times Treated_i + \mu_i + \delta_{it} + \epsilon_{it} \tag{1}$$

where each variable is defined as follows:

- $Y_{it} \in \mathbb{R}$: a welfare measure of individual i in year t, normalized by the base year standard deviation.
- $Treated_i \in \{0, 1\}$: an indicator variable equals 1 if individual i is married as of 2015, 0 otherwise.
- μ_i: individual fixed effects.
- δ_{it}: year-region-education-occupation fixed effects.
- ϵ_{it}: a random error.

and $\mathbb{1}$ is an indicator function. Standard errors are clustered at the individual level.

Individual fixed effects capture individual-level unobserved heterogeneity, and year-region-education-occupation fixed effects capture any macroeconomic shocks specific in a given region in a given education level in a given occupation. Note that occupation is the sector which individual i belongs to, regardless of their employment status, and unaffected by the bill.

I exclude from the sample people below 18 years old and above 74 years old because in Russia,

28. I define one's marital status at 2015 as discussed later, so the effect on unmarried women may also include actual drop in utility. In any case, my estimate on married women is conservative.

29. https://www.hrw.org/news/2016/11/18/russia-thou-shalt-not-disagree-orthodox-church

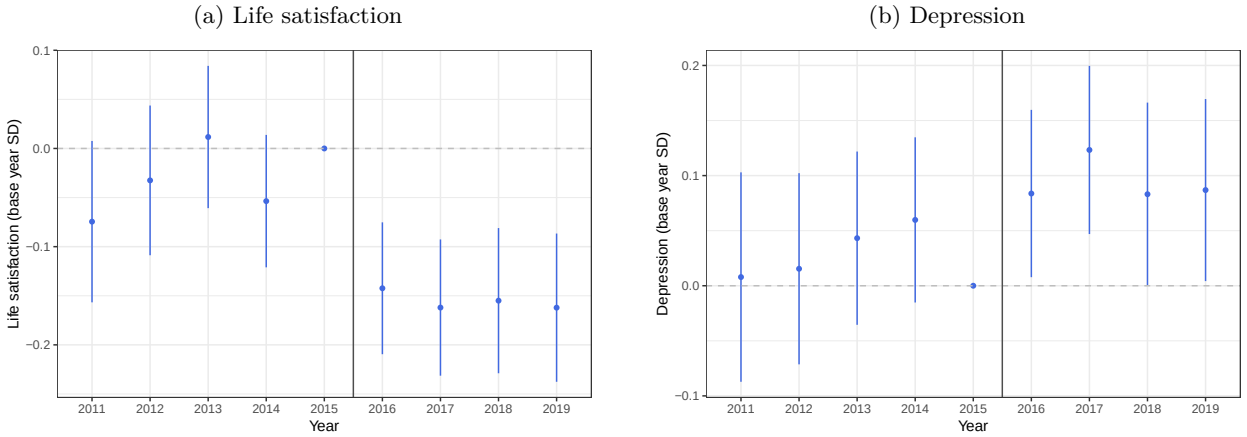
30. As shown in Panel A of Figure 4, this seems a valid assumption: married women's life satisfaction drops from 2016 and stays at the lower level. There is no other event that only affects married women's life satisfaction.

the minimum age to get married is 18, and that spouses of people above 74 years old are likely to be too old to commit domestic violence. I also exclude unmarried women who live with their partners and women who are married but live separately to have a cleaner estimate. Thus, the treated group only includes married women who live with their spouse, and the control group unmarried women who do not live with their partner, both in 2015.

The key identification assumption is the parallel trend: treated and control women's welfare follows the same time trend in the absence of the bill, conditional on time-invariant individual-level unobservables and macroeconomic shocks specific in a given region in a given educational level in a given occupation. Under this assumption, β_ls ($l = 2016, ..., 2019$) capture year-by-year effect of the domestic violence decriminalization bill. β_ls ($l = 2011, ..., 2014$) capture any differential time trend between treated and control women before the bill, which serves as a sanity check of parallel trend assumption.

5 Results

Figure 4: Effect of the bill on married women's welfare



Notes: This figure presents the OLS estimates of β_ls of equation 1 for life satisfaction (Panel A) and depression (Panel B) normalized by the base year (2015) standard deviation along with their 95% confidence intervals. Standard errors are clustered at the individual level.

Figure 4 presents the OLS estimates of β_ls of equation 1 for life satisfaction (Panel A) and depression (Panel B) normalized by the base year (2015) standard deviation along with their 95% confidence intervals. Panel A shows that before the bill, the life satisfaction of the treated and control women roughly follows the same time trend, consistent with the parallel trend assumption. After the bill, however, the treated women's life satisfaction drops and stays at the lower level. Panel B shows that even before the bill, depression levels of the treated and control women follow somewhat different time trends, with treated women trending upward. However, after the bill, the treated women's trend seems to jump up and stays at the higher level.

Table 5 presents standard difference-in-differences estimates from equation 1 to quantify the

Table 5: Effect of the bill on married women's welfare

Dependent variable:	Life satisfaction (base year SD)			Depression in the past 12 months (base year SD)		
	(1)	(2)	(3)	(4)	(5)	(6)
Treated x Post	-0.122***	-0.115***	-0.101***	0.076***	0.062***	0.073***
	(0.017)	(0.019)	(0.020)	(0.019)	(0.020)	(0.021)
Treated x Post x College or above		-0.018			0.036*	
		(0.019)			(0.020)	
Treated x Post x Managers/Professionals			-0.046**			0.008
			(0.018)			(0.020)
Individual FE	✓	✓	✓	✓	✓	✓
Time-Region-Education-Occupation FE	✓	✓	✓	✓	✓	✓
Pre-period mean of treated group	0.633	0.633	0.633	0.072	0.072	0.072
Pre-period SD of treated group	0.238	0.238	0.238	0.259	0.259	0.259
Base year SD of treated group	0.240	0.240	0.240	0.245	0.245	0.245
Adj. R-squared	0.445	0.445	0.445	0.257	0.257	0.257
Observations	51787	51787	51787	51306	51306	51306
Individuals	8961	8961	8961	8942	8942	8942

Notes: This table presents standard difference-in-differences estimates from equation 1. Standard errors in parenthesis are clustered at the individual level. Significance levels: * 10%, ** 5%, and *** 1%.

effect. Column 1 shows that the bill decreased treated women's life satisfaction by 12.2% relative to the base year standard deviation and column 4 shows that the law increased treated women's depression by 7.6% relative to the base year standard deviation. They are quantitatively sizable and statistically highly significant.[31]

The table also shows that the bill decreased treated women's life satisfaction in a highly qualified white-collar occupation more (column 3) and increased the depression level of treated women with a college degree more (column 5). Note that these women are presumably less prone to domestic violence but more sensitive to regressive gender unequal atmosphere.[32] Although statistically insignificant and quantitatively small, the results in columns 2 and 6 are consistent with this story.

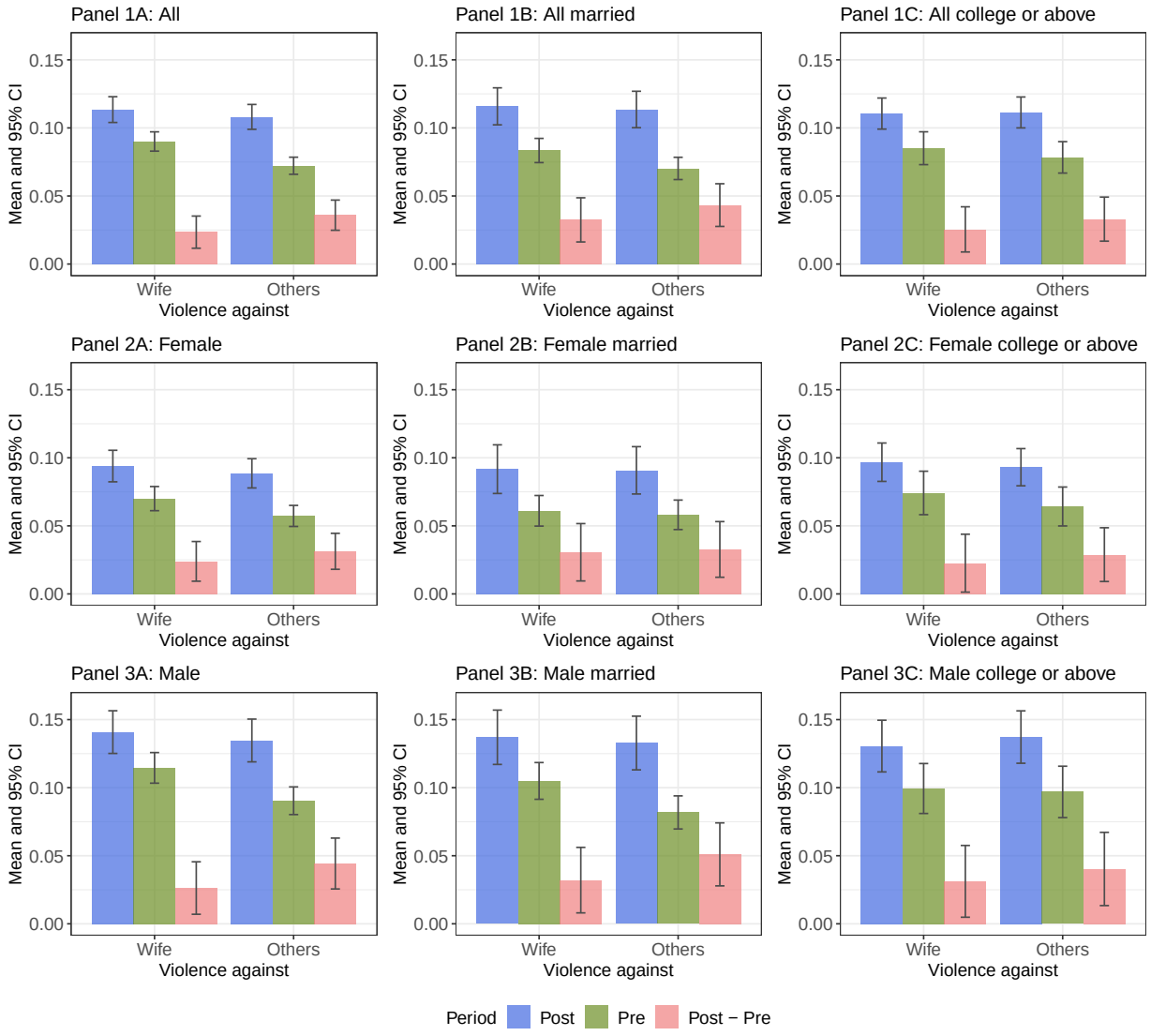
To corroborate the results in Table 5, Figure 5 plots people's attitude toward general and domestic violence before and after the domestic violence decriminalization bill for everyone (Panels 1A-1C), women (Panels 2A-2C), and men (Panels 3A-3C) by their marital status and education level using the WVS data.[33] The figure shows that people became more tolerant toward general as well as domestic violence after the bill, regardless of their gender, marital status, or education level. Thus, these results suggest that the suppressive atmosphere itself may have reduced married women's welfare.

31. Appendix Table A1 gradually adds fixed effects to show the stability of the estimates.

32. As discussed in the introduction, women's economic power has net positive effects in deterring domestic violence.

33. Appendix Figure A3 presents the same plots but for married and highly educated, and shows very similar pictures as Panels 1C, 2C, and 3C.

Figure 5: Changes in people's attitudes toward general and domestic violence



Notes: This figure plots people's attitude toward general and domestic violence before and after the domestic violence decriminalization bill for everyone (Panels 1A-1C), women (Panels 2A-2C), and men (Panels 3A-3C) by their marital status and education level using the WVS data. The 95% confidence intervals are calculated with heteroskedasticity-robust standard errors.

6 Conclusion

This paper studies the effect of increased legal tolerance toward domestic violence on married women's welfare using the Russian domestic violence decriminalization bill introduced to the national congress in 2016. Using difference-in-differences and flexibly controlling for macroeconomic shocks with unmarried women as a control group, I find that the law decreased married women's life satisfaction and increased depression level. The effect is larger for women with a college degree or a highly qualified white-collar occupation who are supposed to be less prone to domestic violence but more

sensitive to regressive gender unequal atmosphere. Consistent with this interpretation, people became more tolerant toward general as well as domestic violence after the bill. These results suggest that the bill reduced women's welfare partly because of the suppressive atmosphere itself.

This paper contributes to studies on the role of legal institutions on domestic violence; this paper provides evidence on how increased legal tolerance toward domestic violence affects married women's welfare in a country with a highly educated female labor force.

References

Aizer, Anna, and Pedro Dal Bó. 2009. "Love, Hate and Murder: Commitment Devices in Violent Relationships." *Journal of Public Economics* 93 (3): 412–428.

American Bar Association. 2014. *Domestic Violence Arrest Policies.* Report. American Bar Association.

Bhalotra, Sonia, Emilia Brito, Damian Clarke, Pilar Larroulet, and Francisco J. Pino. 2021a. *Dynamic Impacts of Lockdown on Domestic Violence: Evidence from Multiple Policy Shifts in Chile.* Working Paper.

Bhalotra, Sonia, Diogo G. C. Britto, Paolo Pinotti, and Breno Sampaio. 2021b. *Job Displacement, Unemployment Benefits and Domestic Violence.* Working Paper.

Boelmann, Barbara, Anna Raute, and Uta Schönberg. 2021. *Wind of Change? Cultural Determinants of Maternal Labor Supply.* Working Paper.

Clerici, Cristina, and Stefano Tripodi. 2021. *Unemployment and Intra-Household Dynamics: The Effect of Male Job Loss on Intimate Partner Violence in Uganda.* Working Paper, Misum Working Paper Series 2021-4. Stockholm School of Economics, Mistra Center for Sustainable Markets (Misum).

Comecon. 2019. In *Encyclopedia Britannica,* by Britannica.

Council of Europe. 2011. *Council of Europe Convention on Preventing and Combating Violence against Women and Domestic Violence (CETS No. 210).* Report. Strasbourg, France: Council of Europe.

Delara, Mahin. 2016. "Mental Health Consequences and Risk Factors of Physical Intimate Partner Violence." *Mental Health in Family Medicine* 12:119–125.

Devries, K. M., J. Y. T. Mak, C. García-Moreno, M. Petzold, J. C. Child, G. Falder, S. Lim, et al. 2013. "The Global Prevalence of Intimate Partner Violence Against Women." *Science* 340 (6140): 1527–1528.

Ericsson, Sanna. 2020. *Backlash: Female Economic Empowerment and Domestic Violence.* Working Paper.

Erten, Bilge, and Pinar Keskin. 2021. "Female Employment and Intimate Partner Violence: Evidence from Syrian Refugee Inflows to Turkey." *Journal of Development Economics* 150:102607.

Federal State Statistics Service. 2017. *Social Status and Standard of Living of the Population of Russia 2017 [Original in Russian].* Moscow, Russia: Federal State Statistics Service.

———. 2021. *Social Status and Standard of Living of the Population of Russia 2021 [Original in Russian].* Moscow, Russia: Federal State Statistics Service.

Garcia-Moreno, Claudia, Henrica AFM Jansen, Mary Ellsberg, Lori Heise, and Charlotte H Watts. 2006. "Prevalence of Intimate Partner Violence: Findings from the WHO Multi-Country Study on Women's Health and Domestic Violence." *The Lancet* 368 (9543): 1260–1269.

Haushofer, Johannes, Charlotte Ringdal, Jeremy P. Shapiro, and Xiao Yu Wang. 2019. *Income Changes and Intimate Partner Violence: Evidence from Unconditional Cash Transfers in Kenya.* Working Paper, Working Paper Series 25627. National Bureau of Economic Research.

Human Rights Watch. n.d. *'I Could Kill You and No One Would Stop Me': Weak State Response to Domestic Violence in Russia.* New York, NY: Human Rights Watch.

Inglehart, Ronald, Christian W. Haerpfer, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Díez-Medrano, Marta Lagos, et al. 2020. *World Values Survey: All Rounds - Country-Pooled Datafile.* Dataset. Madrid, Spain & Vienna, Austria: JD Systems Institute & WVSA Secretariat.

International Labour Office. 2012. *International Standard Classification of Occupations: ISCO-08.* Vol. 1. Geneva, Switzerland: International Labour Organization.

Isajanyan, Nerses. 2017. *Russian Federation: Decriminalization of Domestic Violence.* Report. Washington, DC: The Law Library of Congress, Global Legal Research Center.

Iyengar, Radha. 2009. "Does the Certainty of Arrest Reduce Domestic Violence? Evidence from Mandatory and Recommended Arrest Laws." *Journal of Public Economics* 93 (1): 85–98.

Kozyreva, Polina, Mikhail Kosolapov, and Barry M Popkin. 2016. "Data Resource Profile: The Russia Longitudinal Monitoring Survey–Higher School of Economics (RLMS-HSE) Phase II: Monitoring the Economic and Health Situation in Russia, 1994-2013." *International Journal of Epidemiology* 45 (2): 395–401.

Molina, Teresa, and Mari Tanaka. 2021. "Globalization and Female Empowerment: Evidence from Myanmar." *Economic Development and Cultural Change.*

Monnat, Shannon M., and Raeven Faye Chandler. 2015. "Long Term Physical Health Consequences of Adverse Childhood Experiences." *The Sociological quarterly* 56 (4): 723–752. pmid: 26500379.

Muratori, Caterina. 2021. *The Impact of Abortion Access on Violence Against Women.* Working Paper. University of Turin.

Ridley, Matthew, Gautam Rao, Frank Schilbach, and Vikram Patel. 2020. "Poverty, Depression, and Anxiety: Causal Evidence and Mechanisms." *Science* 370 (6522): eaay0214. pmid: 33303583.

Russian Orthodox Church. 2016. *Statement of the Patriarchal Commission on Family Affairs, Protection of Motherhood and Childhood in Connection with the Adoption of a New Version of Article 116 of the Criminal Code of the Russian Federation [Original in Russian].* Report.

Sanin, Deniz. 2021. *Do Domestic Violence Laws Protect Women From Domestic Violence? Evidence From Rwanda.* Working Paper.

Stevenson, Betsey, and Justin Wolfers. 2006. "Bargaining in the Shadow of the Law: Divorce Laws and Family Distress." *The Quarterly Journal of Economics* 121 (1): 267–288.

Supreme Court of the Russian Federation. 2015. *Plenum of the Supreme Court of the Russian Federation No. 37 [Original in Russian].* Resolution. Moscow, Russia: Supreme Court of the Russian Federation.

The Russian Federation. 1996. *Federal Law of 13.06.1996 No. 63-FZ "Criminal Code of the Russian Federation"[Original in Russian].* The Federal Law.

———. 2016. *Federal Law of 03.07.2016 No. 323-FZ "On Amendments to the Criminal Code of the Russian Federation and the Criminal Procedure Code of the Russian Federation on the Improvement of the Grounds and Procedure for Exemption from Criminal Liability" [Original in Russian].* The Federal Law.

———. 2017. *Federal Law of 07.02.2017 No. 8-FZ "On Amendments to Article 116 of the Criminal Code of the Russian Federation" [Original in Russian].* The Federal Law.

Tur-Prats, Ana. 2019. "Family Types and Intimate Partner Violence: A Historical Perspective." *The Review of Economics and Statistics* 101 (5): 878–891.

United Nations Development Programme. 2020. *Human Development Report 2020.* New York, NY: United Nations Development Programme.

Appendix

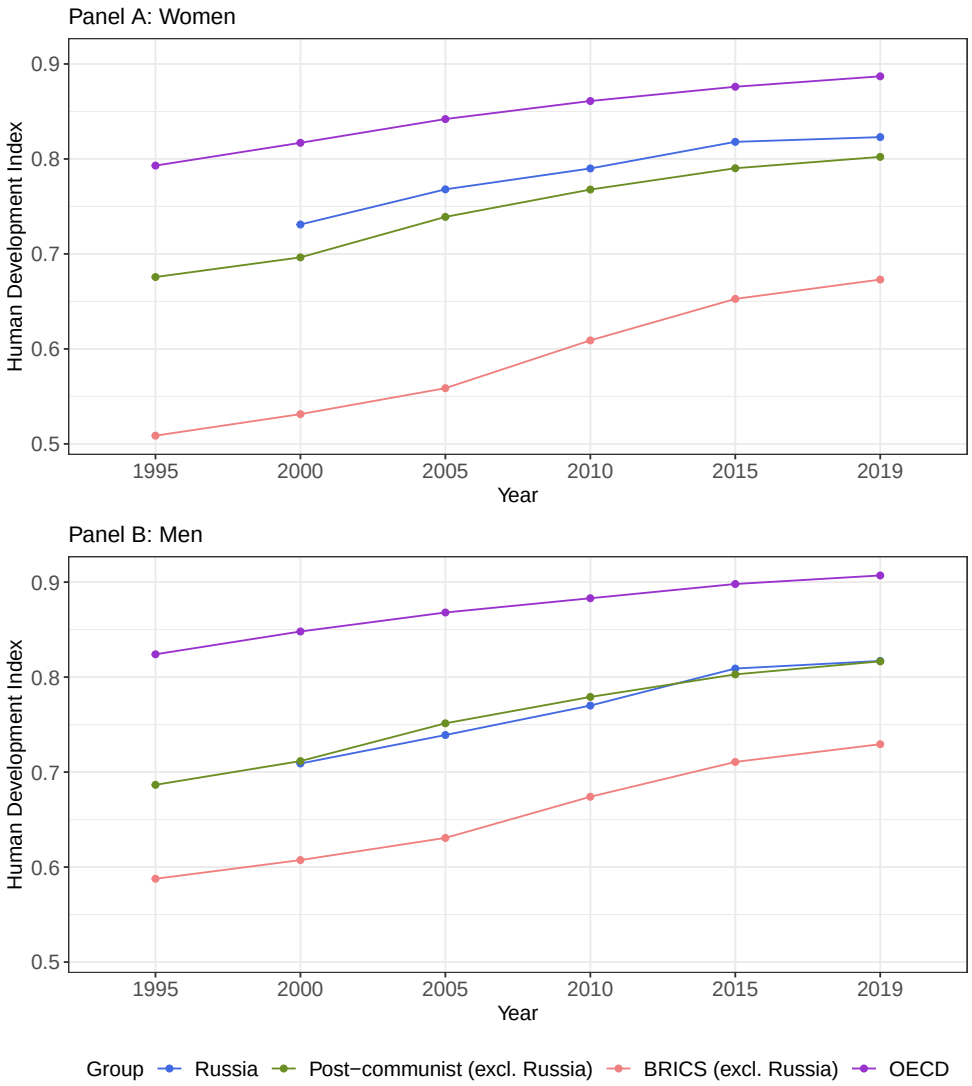
Appendix A Additional Figures and Tables

Table A1: Effect of the bill on married women's welfare (stability of the estimates)

Dependent variable:	Life satisfaction (base year SD)					
	(1)	(2)	(3)	(4)	(5)	(6)
Treated x Post	-0.145***	-0.092***	-0.092***	-0.092***	-0.093***	-0.122***
	(0.014)	(0.013)	(0.013)	(0.013)	(0.013)	(0.017)
Treated	0.531***					
	(0.015)					
Post	-0.091***	-0.020*				
	(0.011)	(0.011)				
Individual FE		✓	✓	✓	✓	✓
Time FE			✓			
Time-Region FE				✓		
Time-Region-Education FE					✓	
Time-Region-Occupation FE						✓
Adj. R-squared	0.050	0.463	0.464	0.464	0.464	0.444
Observations	84812	84812	84812	84812	84668	51853
Individuals	12464	12464	12464	12464	12452	8970
Dependent variable:	Depression in the past 12 months (base year SD)					
	(7)	(8)	(9)	(10)	(11)	(12)
Treated x Post	0.070***	0.051***	0.052***	0.052***	0.052***	0.076***
	(0.015)	(0.015)	(0.015)	(0.015)	(0.015)	(0.019)
Treated	-0.071***					
	(0.014)					
Post	0.078***	0.048***				
	(0.011)	(0.011)				
Individual FE		✓	✓	✓	✓	✓
Time FE			✓			
Time-Region FE				✓		
Time-Region-Education FE					✓	
Time-Region-Occupation FE						✓
Adj. R-squared	0.001	0.290	0.291	0.291	0.291	0.257
Observations	83966	83966	83966	83966	83826	51371
Individuals	12457	12457	12457	12457	12445	8951

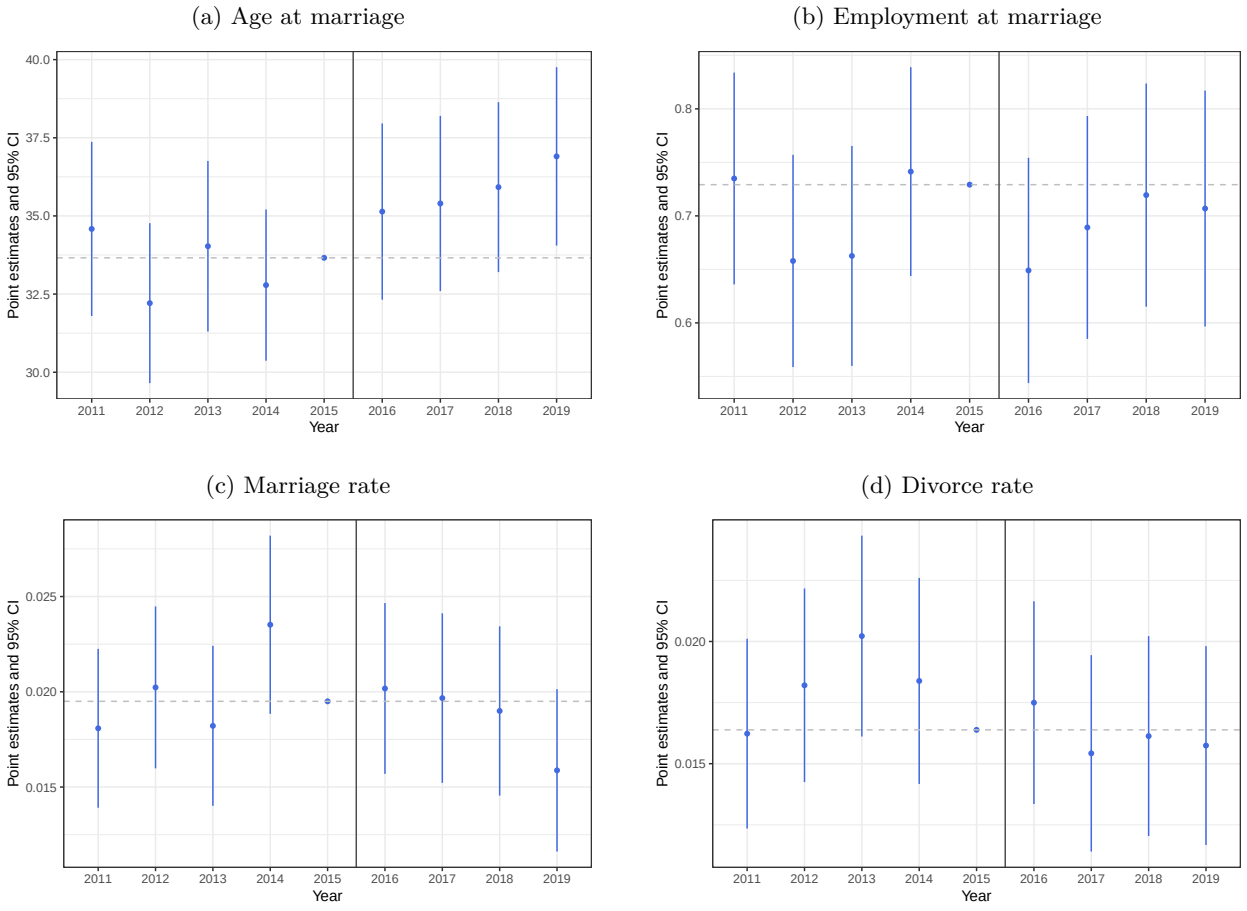
Notes: This table presents standard difference-in-differences estimates from equation 1 but gradually adds fixed effects to show the stability of the estimates. Standard errors in parenthesis are clustered at the individual level. Significance levels: * 10%, ** 5%, and *** 1%.

Figure A1: Human Development Index of Russia and groups of countries, by gender



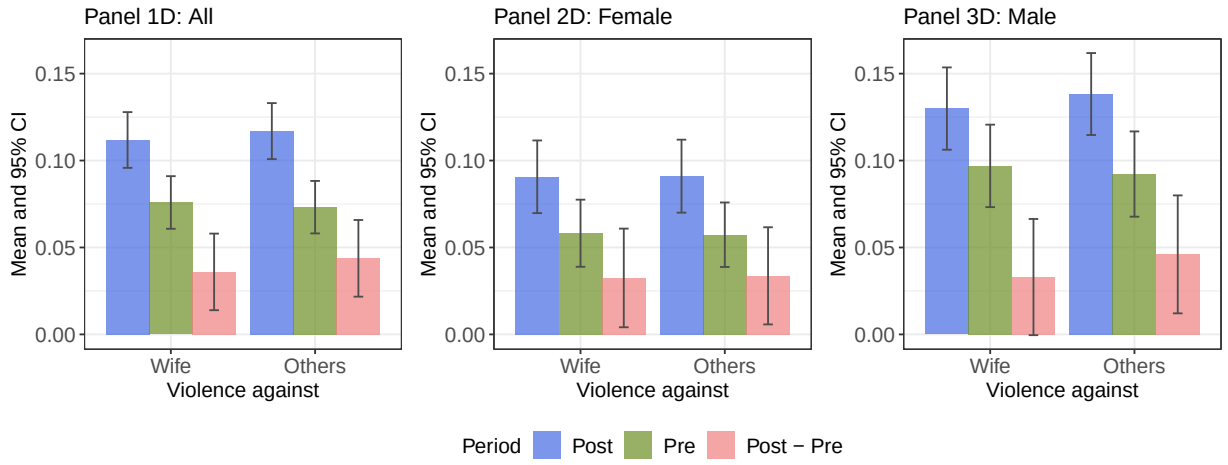
Notes: This figure plots women's (Panel A) and men's (Panel B) human development index (HDI) for Russia (blue), post-communist countries other than Russia (green), BRICS other than Russia (red), and OECD countries (purple) for the period from 1995 to 2019. For the exact calculation of the index, see the technical notes of United Nations Development Programme (2020). BICS are Brazil, India, China, and South Africa. Post-communist countries are initial Comecon members as defined in Britannica (*Comecon* 2019): former Soviet Union (Russia, Armenia, Azerbaijan, Belarus, Estonia, Georgia, Kazakhstan, Kyrgyzstan, Latvia, Lithuania, Moldova, Tajikistan, Turkmenistan, Ukraine, and Uzbekistan), Bulgaria, former Czechoslovakia (Czechia and Slovakia), Hungary, Poland, and Romania.
Source: UNDP Human Development Reports, Gender Development Index (https://hdr.undp.org/en/indicators/136906 for women, https://hdr.undp.org/en/indicators/137006 for men). Retrieved on February 15, 2022.

Figure A2: Women's selection into marriage and divorce



Notes: This figure presents simple year-by-year average and 95% confidence intervals of age at marriage (Panel A), employment status at marriage (Panel B), marriage rate (Panel C), and divorce rate (Panel D) around the bill, relative to the base year (2015). The confidence intervals are heteroskedasticity-robust.

Figure A3: Changes in people's attitudes toward general and domestic violence: Married and college or above



Notes: This figure presents the same plots as Figure 5 but for married and highly educated. The 95% confidence intervals are calculated with heteroskedasticity-robust standard errors.