

# It Takes Two: Gender Differences in Group Work\*

Siri Isaksson <sup>†</sup>

November 20, 2018

## Abstract

This study tests for gender differences in credit claimed for individual contributions to group work. I introduce a novel experimental design in which two subjects work together on solving a computerized puzzle, by making alternating moves. Participants play nine rounds, each time with a new partner and puzzle. After each puzzle, they are asked to estimate their contributions towards the solution in incentivized questions. There are no gender differences in ability: women and men are equally good at solving the puzzle both individually and in teams. Despite their equal contribution, women consistently claim less credit than men. This effect is strongest among high contributing women, and women in groups that implemented more complex solutions. I also explore the propensity of participants to undo a partner's move, and I find that men are more likely to correct a partner when he or she made a move that was wrong. These results suggest that gender differences in claiming credit may contribute to the labor market gender gap.

---

\*I would like to thank my advisors Magnus Johannesson, Katherine Baldiga Coffman and Anna Dreber Almenberg for their valuable feedback and support. I thank Adam Sam at Semsomi, who developed the software used in this study, and supported the data collection. I thank Laura Adler, Adam Altmejd, Iris Bohnet, Hannah Riley Bowles, Yiling Chen, Zoe Cullen, Tore Ellingsen, Clémentine van Effenterre, Ben Enke, Christine Exley, Lily Hu, Alex Imas, and David Parkes for helpful advice. I thank the Women and Public Policy Program at Harvard Kennedy School and the Econ CS group at Harvard SEAS where I have been a visiting fellow while working on this project. I thank seminar participants at the North American ESA conference 2017 and 2018, the ESA world conference 2018, the WAPPP seminar series, the 2018 SITE summer workshop, the Berlin Behavioral Economics Seminar, the Choice Lab at NHH, the UC San Diego Spring School in Behavioral Economics and the Stockholm School of Economics Brown Bag seminar for helpful comments. Financial support from the Tom Hedelius and Wallander Foundation, Vetenskapsrådet, and the Sweden America Foundation is gratefully acknowledged. Finally, I would like to thank the staff at HDSL for their assistance with the data collection, in particular Alki Iliopoulou for help with recruitment. This study was approved by the IRB (IRB18-1138). All remaining errors are my own.

<sup>†</sup>Stockholm School of Economics, Harvard Kennedy School. *Corresponding address:* 67 JFK street, Cambridge 02138 MA, USA, *Email:* siri.isaksson@hks.harvard.edu *Phone:* +1 857 919 9627

# Introduction

Regardless of which professional career we choose in life, it is likely to involve teamwork: researchers co-author papers, consultants collaborate to build client solutions, programmers revise one another's code, and sales teams work together to reach their monthly quotas. In fact, team work increased from 27 to 87 percent as the share of work completed in large US firms between 1980 and 1996, and remains at high levels (see Lazear and Shaw 2007).

One distinguishing characteristic of group work is that individual contributions are often ambiguous. This raises concerns about whether credit will be allocated fairly, and potentially creates room for bias to flourish. The present paper asks whether there are gender differences in claiming credit following successful group work. That is, conditional on making the same contribution to group success, do women claim less credit for themselves?

I explore gender differences in group work using a computerized laboratory experiment where subjects ( $N=197$ ) first solve puzzles individually and then in teams of two. Specifically, I designed a game in which I randomly assign participants to pairs who solve a 3x3 sliding puzzle together by taking turns in making a move. If a participant thinks that her partner made a mistake, she can correct her partner. The mathematical properties of this puzzle make it possible to precisely track whether each move is good, or bad, permitting a clear measure of individual contributions to shared success. In addition, it is also possible to understand whether participants are right to correct their partners in this setting. The experimental design, hypotheses, and analysis were pre-registered.

There are two important benefits of studying group work in the laboratory that motivated using this setting. First, the laboratory allows me to randomize team members into groups whereas in many field settings people select into teams, for instance because they like someone, or because they know that someone is productive. This makes it hard to identify clearly the role of gender in naturally formed groups. By randomly assigning team members to groups, I solve this endogeneity issue. A second challenge in studying group work in the field is measuring and comparing different types of contributions. By instead using the laboratory, I can create a quantifiable, clear,

measurable definition of contribution and use it to compare participants of identical contributions. This allows for consistent comparison.

I have four main findings. First, there are no gender differences in ability to solve the puzzle introduced in this paper. When working individually, women solve 7.4 puzzles, whereas men solve 6.9, this difference is not statistically significant. When working in pairs, women and men contribute equally to the solution on average. Second: despite this, I find that women under-credit their contribution by 4.4 percentage points. Given the fact that average contribution for both men and women is 50%, this effect size corresponds to a 8.8% gender gap in claiming appropriate credit for contributions to group work. Third, this effect is strongest among high contributing women, who under-credit their contribution by 5.1 percentage points, and women who work on a complex solution who under-credit their contribution by 5.5 percentage points. Fourth, I measure the tendency to correct one's partner's previous move, finding that men are 3.5 percentage points more likely to correct a partner's mistake. This corresponds to a 31% gender gap in the propensity to correct a partner who just made a mistake. These results are robust when controlling for various characteristics of the game, for instance how much time it took to solve a puzzle.

Given the prevalence of collaborative group work, research on how women and men interact and value themselves in teams is necessary to understand gender inequalities in the labor market (Azmat and Petrongolo, 2014). Yet, experimental studies on gender differences in team interactions are relatively scarce. I define team work as work done in a group consisting of two people or more, who engage interactively to achieve a common goal. A small but burgeoning literature on gender and group work finds that expert women are less likely to contribute ideas to a group Coffman (2014) and that women are less willing to lead male dominated groups Born, Ranehill and Sandberg (2018). Others have examined how gender composition affects performance Ivanova-Stenzel and Kübler (2011) and Hoogendoorn, Oosterbeek and Van Praag (2013). Focusing on individual decision making, research in this field has established that women are less confident and less likely to enter competitions than men (see Niederle and Vesterlund (2011) for a survey), and that women on average are more risk-averse than men (see Croson and Gneezy (2009) for a

survey). I build on this work by investigating co-ordination, co-operation, and credit claiming in a carefully controlled environment. In addition, I make an important methodological contribution by introducing a new task with several attractive features, such as being able to precisely track individual contributions to shared work. This task can be used in future studies to examine different aspects of group work along any demographic axis, and could easily be adapted to study the dynamics in groups consisting of three or more members.

To summarize, even though women are as good as men at this task, both individually and as part of a team, I find that women under-credit their contribution following successful group work. Systematically undervaluing contributions to shared work may lead to gender differences in lifetime labor market outcomes. Recent research shows that despite advances for gender equality in the labor market, women remain disadvantaged both in terms of hiring, wage and promotions, especially at the top levels of companies (see Blau and Kahn 2017 for a survey). In this paper, I suggest that gender differences in claiming credit might be one behavioral channel through which the labor market gender gap persists.

## 1 Introducing the Puzzle

I explore gender differences in group work through a laboratory experiment in which participants solve different puzzles first on their own, and then together with different partners in exogenously formed pairs. This requires a task fulfilling the following four criteria:

1. The task has to involve interaction within the group, allowing participants to work together towards a common goal. This is important because one of the defining characteristics of group work is that the outcome is something created together, rather than simply the sum of two individual inputs.
2. Individual contributions should not be immediately clear to participants, so a potential bias in self-attribution of credit has the possibility to flourish. As discussed by Haynes and Heilman (2013), ambiguity in individual contributions to shared work allows gender bias in attribution

of credit to grow. This makes sense since if individual contributions are immediately clear, claiming and attributing appropriate credit for a joint success is an easier task than if they are obscured.

3. It is crucial that individual contributions can be clearly and objectively measured by the researcher in order to establish whether participants are claiming appropriate credit or not. Arguably, gender differences in claiming credit are mainly interesting if we can back out actual contribution levels. If, for instance, men would claim and make higher contributions, their claims would be appropriate. Thus in order to understand whether there are gender differences in claiming appropriate credit, we need to be able to control for actual contribution levels. In fact, as argued by for instance Anderson et al. (2012), the most important aspect in measuring any type of overly positive self evaluation is being able to objectively back out the true individual performance from the perceived.

With these criteria in mind, I create a new task to answer my research questions. Specifically, I design a game in which participants solve different 3x3 sliding puzzles on a computer in teams of two. When a pair starts solving the puzzle, the tiles are in the wrong order, see figure 1 panel a which shows the screen of a participant working together with Maria. Maria is a fictional player used to illustrate how the puzzle works. However in the experiment, participants would see the real first name and picture (as captured in the laboratory) of their partner on their screen. In order to make sure participants knew whether they were working with a female or a male partner, I use both the picture and the first name to introduce participants to each other, since first names are often gender ambiguous. Teammates sit in different rooms, follow each other's moves on the screen, and have no way of communicating other than through their moves. The puzzle is solved when all tiles are in numeric order, leaving the right bottom corner empty (see figure 1 panel c). In order to solve a puzzle, a sufficient number of good moves needs to be made, net of bad moves. A participant's individual contribution to the solution is defined as her number of good net bad moves relative to those made by her partner, this is described in more detail below. Due to the fact that teammates cannot communicate, co-operation and co-ordination among team members

is needed to solve the puzzle. Simply put, it takes two to solve this puzzle. No matter how good one player is on her own, she has to rely on her partner in order to be successful. This makes sure that the first criterion above is fulfilled.

Participants have four minutes to solve the puzzle. A clock in the top right corner keeps track of how much time is remaining. In figure 1 panel a, the participant is waiting for Maria to make her move. Maria has three options: she can either move 7 down, 4 right or 6 left. If Maria moves 4 right, the board will now look like the one seen in panel b of figure 1. At this point, it will be the other participant's turn to make her move. She now has two possible moves to choose from: she can either move 8 down or 4 back left. I focus on successful group work, because it often plays a part in decisions of who gets awarded a promotion, tenure, bonuses and other important labor market outcomes. The experiment was thus designed to study gender differences in claiming credit for a joint success, and it was pre-specified in the pre-analysis plan to only use credit claims from solved puzzles. Therefore, it was crucial that a majority of puzzles were solved. To help participants along the way, the first row is always sorted in advance, as seen in figure 1 panel a - c, and the tiles 1, 2, and 3 could not be moved.

A tile is moved by clicking on it once. A tile can be moved into the adjacent empty space either vertically or horizontally, but not diagonally. Each move is either good in the sense that it moves the team closer to the solution, or bad in the sense that it moves the team further away from the solution. Good and bad moves are determined using a Breadth First Search algorithm, described in more detail later in this section. There are no neutral moves. For example, in figure 1 panel a, moving 7 down is good, and moving 4 right is bad.<sup>1</sup>

Whether a move is good or bad is determined by a Breadth First Search (BFS) algorithm (see Bundy and Wallen (1984)). For each configuration of tiles, BFS finds and logs the minimum number of moves needed to solve the puzzle. For instance, the BFS solution to the puzzle seen in figure 1 panel a is seven, since this puzzle needs at least seven moves to be solved. Each move

---

<sup>1</sup>Note that depending on what configuration a player is in when she decides on her move, there can be either two or three alternatives to choose from. For instance, in the configuration seen in figure 1 panel b, two moves can be made whereas in the one in figure 1 panel a, there are three possible moves. It follows that while all moves are either good or bad, they are not uniquely good or bad. Given a configuration, there can be multiple good, or multiple bad moves.

creates a new configuration of tiles associated with a new BFS solution.

Following a move, the fastest path to solution is either increased or decreased by one move. Returning to figure 1 panel a, a configuration for which the fastest path requires seven moves, if Maria moves tile 7 down, the new BFS solution will be six: through her move she has reduced the shortest path to solution from seven to six moves. This is counted as a good move. If instead she chooses to move 4 right, she would increase the shortest path to solution from seven to eight moves. This is considered a bad move. Participants are not told whether their, or their partner's moves are good or bad. Since a BFS solver is needed to determine this, participants have to rely on their subjective judgment to decide whether a move is good or bad. This feature of the puzzle ensures that individual contributions are ambiguous and that the second criterion is fulfilled. If a participant thinks that her partner made a bad move, she can reverse that move.<sup>2</sup> For example, going back to figure 1 panel a, if Maria chooses to move 4 right, and her partner believes that this was a bad move, she can move 4 back to the left. A move that reverses a partner's move in this manner, is called a reversion in the remainder of this paper. Since there are no neutral moves, it follows that there are no neutral reversions. If a player reverts a bad move made by her partner, I count this as a correction; if she reverts a good move, this is considered a bad reversion. For each move, including the reversions, the system logs the move in an individual moves vector with a +1 if it is good move and -1 if it is bad move.

When the puzzle is solved, or the time is up, I calculate a score equal to the sum of all good net bad moves for each player. A player's relative score is her contribution. For example, say the puzzle in figure 1 panel a is solved in 9 moves by the two players Maria, who first does 2 bad and then 3 good moves, and her partner, who does 4 good and no bad moves. In this example, Maria's moves vector would be  $[-1, -1, 1, 1, 1]$ . Her score would be 1, the sum over her moves vector. Her partner's moves vector would be  $[1, 1, 1, 1]$ , and her score would be 4. Maria would have contributed 20% ( $1/(1+4)$ ), and her partner the remaining 80% ( $4/(1+4)$ ). In the case that a puzzle was solved, but one of the players made more bad than good moves, I cap that

---

<sup>2</sup>Note that it follows from the fact that reversion is possible that one move can never increase or decrease the fastest path by more than one move.

player's contribution at 0% to make sure that contributions are always in the closed set of 0% to 100%. In such cases, one player contributed 0% and the other contributed 100%. Thus, using the mathematical properties of this puzzle allows me to precisely track individual contributions to shared work, which accounts for the third criterion above.

As participants take turns in making their moves, they are not told whether the moves they make are good or bad. They are however aware of the fact that their moves are either good or bad, and how to calculate their contribution using these concepts. By construction, the individual contribution ranges from 0% to 100% depending on how many good net bad moves a player made. Individual contributions within each team will always sum to 100% if the puzzle is solved. After a puzzle is solved or the time is up, each participant is asked with incentives to estimate their contribution (see figure 2). The question is incentivized such that participants earn more money the closer their credit claim is to their true contribution. The answer to the credit question ranges between 0% and 100% and is the main outcome variable studied in this paper.<sup>3</sup>

After both participants have answered the credit question, they are matched with a new puzzle and a new partner and the task is repeated. By new puzzle, I am referring to a new starting configuration. As described above, each configuration of tiles is associated with a value equal to the minimum number of moves needed to solve the board. Thus, different starting configurations correspond to different difficulty levels. For example, the easiest puzzle started at least 2 moves from the solution and the hardest started at least 19 moves from solution. By varying the starting configuration of the tiles for each puzzle, I can collect several observations from the same task, across varying levels of difficulty without tedious repetition.<sup>4</sup>

---

<sup>3</sup>I use the binarized scoring rule from Hossain and Okui (2013) to incentivize the credit question. The binarized scoring rule was selected because it is neutral to different risk preferences, which frequently differ between men and women.

<sup>4</sup>Initially, I wanted to randomize starting configurations, however since some configurations unsolvable, I instead pretested 26 starting configurations which I knew were solvable, see Archer (1999) for a discussion.



## 2 Experimental Design

I performed a computerized laboratory experiment at Harvard Decision Science Laboratory in the Fall semester 2018. I collected data from  $N = 197$  (109 male and 88 female) participants.<sup>5</sup> To be eligible, participants had to be either currently enrolled in school, or have some sort of degree. This included everyone with: a high school diploma/GED, some college (not received degree), associates degree, professional certification and/or license, bachelor's degree, some graduate school (not received degree), graduate or professional degree, current full time college student, current full time graduate student, current part time college student and current part time graduate student. In addition, everyone who was recruited self-identified as either male or female. The randomization occurred on four levels. First, the seating was randomized: participants were assigned a seat by the researcher and were not allowed to select their seats on their own. Second, the team assignments were randomized such that gender composition was a random variable. Third, the sequence of puzzles was randomized. Fourth, the opportunity to make the first move was randomized.

In order to study the role of gender in group work, each session had to fulfill two criteria. First, there had to be enough women and men within each session so that I could match subjects into each possible gender composition: mixed teams, female teams and male teams. Since the participants in one room are paired with participants in the adjacent room into teams, both genders had to be present in both rooms, optimally in equal proportions. This was addressed by the recruitment strategy. The second criteria for being able to study the role of gender in group work, is that participants are randomized into pairs so that the gender composition of a group becomes a random variable. This criterion was fulfilled by the way I seated participants, and the way I matched participants into pairs. Thus the recruitment, seating, and matching of participants are all crucial to the identification strategy of this experiment. Below, I go through in detail how each of these were designed and executed.

Starting with the recruitment, first note that I used two rooms, room B and room C, in the HDSL, each of which has 12 computers. So the maximum capacity was 24 subjects in a

---

<sup>5</sup>The experimental design was pre-registered together with the hypotheses and statistical tests on OSF. Please view the following link to view this pre-analysis plan: <https://osf.io/bskdm/>

session. Recall that the aim of the recruitment was to have an equal number of men and women in both rooms, so I tried to recruit 12 men and 12 women to each session. The actual number of participants was dependent on how many subjects showed up to a session, and ranged from 10 to 18 across sessions.

Each participant was seated by the researcher in an individual computer cubicle. The way the matching works is that in the first round, participants with the same seat digit work together, so B1 works with C1, B2 with C2, and so forth. Then, in the next round, the B's rotate, so that — if the seats are filled up until for example seat 8 — C1 now works with B8, C2 with B1, C3 with B2 and so forth until C8 who works with B7. The rotation of B seats continues in this manner until every possible match has been made. There had to be two additional subjects seated on two additional seats, one in each room, for each additional puzzle added to the sequence. So if 14 participants came to one session, C1-C7 and B1-B7 were all seated, and 7 puzzles were included in that session. If 10 subjects showed up, C1-C5 and B1-B5 were seated, and 5 puzzles were included. The length of a session thus varied, as each puzzle added 4 minutes to the session. Sessions lasted between 70 and 100 minutes.

Having given consent, participants read the instructions on their screens, which explained how the puzzle works, how contributions are defined, and how they would be compensated. The instructions urged participants to try their hardest to solve every puzzle and to answer questions correctly. The instructions also included a slide-show example of a puzzle being solved, move by move. For each move showed in the slide-show, participants were told if it was good or bad. They were then asked to calculate each example participant's contribution. This was done to ensure that the concept of contribution was clear. Most participants answered these questions without problems. If a participant didn't know how to answer one of the questions, the researcher provided a verbal explanation of the reasoning necessary to answer the questions.

After a participant had successfully answered all questions, she was automatically redirected to a sign in screen (see figure 11). Here, the participant was asked to provide her first name and seat number and take a photo of herself using the camera attached to computer. At the bottom of the

sign-in screen was a box for the researcher’s signature. This gave the researcher the opportunity to check that every participant had entered the correct first name, seat number, and that the photo was neutral. If a photo stood out for some reason, for instance because a participant had a cap on, the researcher asked the participant to remove the cap, and a new picture was taken. This matters since the goal of the photo was to signal gender, and not e.g favorite baseball team.<sup>6</sup>

Once all information on the screen was correct, the researcher signed in the participant and the practice round started. In the practice round, participants had 4 minutes to solve as many practice puzzles as possible on their own. In order to make sure that participants tried their hardest, the practice puzzles were incentivized at a rate of 25 cents per solved puzzle (see figure 12). There were 15 practice puzzles and the measure of individual performance is the number of practice boards solved, ranging from 0 to 15. The practice boards were selected from 26 puzzles that had been pretested in the laboratory. There were two goals when designing the practice round. First, practice would help participants achieve a higher rate of success in their teams.<sup>7</sup> Secondly, the practice round allowed me to measure individual performance in this puzzle. To this end, puzzles ranged in difficulty, see figure 12 which shows the easiest practice puzzle.

After participants finished their practice round, they were shown a list of the names and pictures of all the people in the adjacent room, and were asked, with incentives, to guess how much they thought each of their future partners would contribute on average, ranging from 0% to 100% (see figure 13).<sup>8</sup> Below is an excerpt of the text participants read on their screen when they were prompted to rate their future partners:

*Below is a list of all the people that you will be solving the sliding puzzle with today. Each person is presented together with their name and picture. Each partner that you will be matched with will try to solve the puzzle with everyone in your room. So just like you, each person will have*

---

<sup>6</sup>As an additional effort to neutralize pictures, angle and zoom of the camera were standardized. In a follow-up study, pictures will also be rated on various aspects of appearance to be able to control for e.g the effect of appearing attractive for credit claims.

<sup>7</sup>This was important since I only use credit data from solved puzzles, so I wanted to make sure that a majority of puzzles attempted in teams would be solved.

<sup>8</sup>Just as with the credit question, we use the binarized scoring rule from Hossain and Okui (2013) to incentivize this question.

*tried  $X^9$  puzzles in total by the end of the experiment. Please make an estimate ranging from 0% to 100% of how much you think each person's average contribution in these  $X$  puzzles will be.*

The answer to this question was used to determine whether the task was perceived as male- or female-stereotyped. This measure can also be used as a proxy for gender differences in individual self-confidence, since a high answer on this question indicates low self-confidence as it attributes a higher expected contribution to the partner than to the self.<sup>10</sup> Once a person had answered the prior question for each future partner, she was directed to the first puzzle with an assigned partner and it was randomly determined who would make the first move.

Once participants had solved the puzzle or four minutes had passed, a question appeared on the screen about how much each participant thought they had contributed to the solution (see figure 2). The text explained that the answer would not be revealed to the partner. The participants gave an answer between 0% and 100%. The correct answer was the sum of good minus bad moves that a participant had made in a game relative to her partner. Below the credit question, participants were prompted to answer another question *"Do you think you would be better on your own?"*. This question aimed to measure subjective sentiment. The five answer alternative were: *"Yes, I would have solved it much better on my own."*, *"Yes, I think I would have solved it somewhat better on my own."*, *"No, I think it would have stayed the same."*, *"No, I think I would have solved it somewhat worse on my own."*, and *"No, I think I would have solved it much worse on my own."*. Once a subject had answered these questions, they were matched with a new partner and a new puzzle according to the matching mechanism previously explained. The task was repeated until every person in one room has attempted a puzzle with every person in the other room.

Finally, participants were redirected to a page where they were asked demographic questions about their gender, age, field of study, place of birth, and ethnicity. Each participant also completed a post-study survey asking broad questions about the experiment: *"What do you think this experiment was about?"*, *"Was there anything unclear or confusing about this study?"* *"Do you*

---

<sup>9</sup>This number ranges from 5 to 9 puzzles.

<sup>10</sup>The interpretation on self-confidence follows from the fact that both players' contributions in a solved puzzle always sum up to 100%.

*think the puzzles were easy, ok, hard or very hard?*, *Do you have any comments on the puzzles?* and *"Do you have any other comments?"* For each subject, one out of all the puzzles attempted in pairs was randomly selected for payment. If that puzzle was solved, the participant for whom that puzzle was selected earned an additional \$10, if it was not, she earned 0\$ for her group work. I chose to randomly select one of the team puzzles for payment to ensure that participants would try their hardest on every single puzzle. In addition, each solved puzzle in the individual round generated an additional 25 cents.<sup>11</sup> Since there were 15 practice puzzles, participants could earn between \$0 and \$3.75 from this round depending on how well they did. Finally, one of the questions on prior beliefs and one credit question was randomly selected, and if the participant provided an answer close enough to the actual contributions on these questions, she earned an additional \$2 per question. The binarized scoring rule from Hossain and Okui (2013) was used to determine how close claims had to be to the objective contribution in order for the subject to earn the additional \$2 on these questions. This scoring rule was used as it is incentive compatible for different risk preferences, and gender and risk preferences often interact.<sup>12</sup> Finally, everyone earned the show up fee of \$15. Thus, participants could earn maximum \$32.75 and were guaranteed \$15.

### 3 Relationship to Experiment 1

The puzzle introduced in this paper was used in a previous experiment in the Spring semester of 2017 when data from 248 students was collected at the HDSL. In the following, I refer to the previous data collection as Experiment 1 and the current one as Experiment 2. The insights from Experiment 1 informed Experiment 2 in two ways. First, the design in Experiment 2 was updated using information from Experiment 1. For example, the post study survey in Experiment 1 made clear that participants thought that the first move was important for team success. As a consequence, the first mover within each pair was randomized in Experiment 2. The improvements in design between Experiment 1 and 2 were substantial and are described in detail in a working

---

<sup>11</sup>Note that randomization for payment would not have been incentive compatible in the individual round, since the payoff maximizing strategy would then have been to only solve one puzzle in the practice round.

<sup>12</sup>See Croson and Gneezy (2009) for a discussion

paper which is available upon request. Second, the data collected in Experiment 1 was used to help formulate the research questions for Experiment 2, and the hypothesized direction for each test. For example, the data from Experiment 1 indicated that men had a higher propensity to make corrections than women. This informed the hypothesized direction for the test on whether gender affects the propensity to make corrections. Also note that the data collected in Experiment 2 is richer than the data from Experiment 1. For instance, Experiment 1 did not include an individual practice round, and thus there is no way to control for individual ability. Due to the fact that the data in Experiment 2 is richer and the experimental design was improved in several ways, the datasets can not be pooled, the results are not immediately comparable, and the data from Experiment 2 is more informative. The four research questions for Experiment 2, and their associated hypothesized direction, based Experiment 1 are outlined below.

1. Does team gender composition matter for team success? — **Direction:** Yes, mixed teams are less likely to solve a puzzle than male teams.
2. Does team gender composition affect how well participants perform individually? — **Direction:** Yes, performance is lowered when working with a female partner.
3. Do women under-credit their contribution to group work? — **Direction:** Yes, women under-credit their contribution to group work.
4. Are men and women equally likely to reverse a partner’s move? Are they equally likely to have their moves reversed? — **Direction:** Women make and are the target of less corrections than male participants. Turning to bad reversions, men make more of these, while women are more likely to be the target.

The first and second hypothesis were based on suggestive evidence in Experiment 1, indicating that mixed teams were less successful than male teams and that male players lowered their individual performance when working together with women in teams. However, in Experiment 1, I did not record the same clear measure of individual performance as I do in Experiment 2.<sup>13</sup>

---

<sup>13</sup>Specifically, in Experiment 1, a truncated measure for individual performance was used, namely the number of

The third hypothesis was based on the fact that I saw a pattern of women under-crediting their contributions in the first round in Experiment 1, especially when working towards a complex solution. In Experiment 1, the sequence of boards was not randomized, instead I used one complex and one simple sequence of puzzles. Because the aim was to study gender differences in successful group work, it was important to have as many solved puzzles as possible. Since Experiment 1 did not have a practice round — due to practical constraints — I was reluctant to randomize the sequence of puzzles solved in pairs as my intuition was that this would lower the solution rate substantially. Moreover, Experiment 1 involved a within subjects treatment variation that was removed in Experiment 2. To summarize, the main improvements between Experiment 1 and 2: an individual practice round was added, two levels of randomization were added, and a within subjects treatment variation was removed. I hypothesized that with this improved design, the noise observed in Experiment 1 would be decreased, and women would significantly under-credit their contribution to shared work. The fourth hypothesis was similarly based on a pattern in the data in Experiment 1 indicating that gender matters for how efficiently participants reverse each other’s moves.

## 4 Empirical Strategy

The hypotheses in this paper were inspired and informed by previous research showing that women are less confident, and less likely to enter competitions than men (see Niederle and Vesterlund (2011) for a survey). Based on these findings, my prior was that women would under-credit their contributions to a shared success. Confidence also plays into the decision to undo a partner’s move, since it requires being sure enough in the own move to undo the partner’s while that partner is watching. Thus, I hypothesized that women would be less likely to undo a partner’s move.<sup>14</sup> The hypothesized direction for each test, and the research questions were also informed by Experiment 1 described in section 3. In the present section, I go over the four regression models that were used

---

good moves minus the bad ones truncated at zero. Experiment 2 instead uses an untruncated measure of individual performance: the untruncated sum over the moves vector. Thus, the measure in Experiment 2 is more informative.

<sup>14</sup>For more detail, please refer my pre-analysis plan: <https://osf.io/bskdm/>.

to test the four research questions in this paper.<sup>15</sup>

### Question 1 — Does team gender composition matter for team success?

$$complete_{i,j} = MixedTeam_{i,j} + FemaleTeam_{i,j} + X_{i,j} + \epsilon_{i,j} \quad (1)$$

The outcome variable  $complete_{i,j}$  is a dummy for the game that individual  $i$  attempts in round  $j$  equal to one if the solution was reached. Note that every test in this section uses the same set of controls, denoted as  $X$  in the regression models.<sup>16</sup> I use male teams as a baseline and check whether mixed teams or female teams have a lower probability of reaching the solution.  $MixedTeam_{i,j}$  is a dummy equal to one if the team that individual  $i$  belongs to in round  $j$  has both a female and a male team member. The coefficient shows the percentage point change in probability of success in mixed teams compared to male teams.  $FemaleTeam_{i,j}$  is a dummy equal to one if player  $i$  is female and her partner in round  $j$  is also female. The coefficient shows the percentage point change in probability of success of a female team compared to a male team. I use two-way clustering and cluster by participant and team.<sup>17</sup> This model allows me to test for how gender composition impacts group success. Previous literature has offered divergent findings on how gender composition of a team affects team success, see for instance Hoogendoorn, Oosterbeek and Van Praag (2013) and Ivanova-Stenzel and Kübler (2011). In this setting, using the insights from Experiment 1, I hypothesized that male teams would have a higher success rate than mixed teams.

---

<sup>15</sup>Each of the tests are two-sided and use 5 percent as the significance level. Unless otherwise noted, I use STATA's `ivreg2` command in these regressions, in order to be able to use two-way clustering. This is needed to address the fact that each observation occurs twice in the data, both as a player and as a partner.

<sup>16</sup>The set of controls for each test were pre-specified in the preanalysis plan are divided into controls for the player  $i$  and the partner  $i$  meets in round  $j$ . The controls are: individual ability (an integer between 0 - 15 measuring the number of puzzles that the participant solved in the practice round), age, where the participant grew up, ethnicity, and level of education. Dummies are used for categorical controls. Information on field of study is not used due to selection on gender into field.

<sup>17</sup>The clustering on the team level means that each pairing of two people (which is unique) has one cluster. Since each pair occurs twice in the data, both for the player and for the partner, this variable was used here for clustering.



**Question 2 — Does team gender composition affect how well participants perform individually?**

$$performance_{i,j} = Female\ Player_i + Female\ Partner_{i,j} + X_{i,j} + \epsilon_{i,j} \quad (2)$$

The outcome variable  $performance_{i,j}$  is the net sum of good moves minus bad moves that each player  $i$  makes in each round  $j$ . The coefficient on  $Female\ Partner_{i,j}$  measures the change in individual performance when a participant works with a female as opposed to a male. I also run a separate regression adding a dummy for female teams to the model to examine gender differences in the effect of having a female partner. The coefficient on the female team dummy measures gender differences in the effect of having a female partner on individual performance. I cluster by participant and partner. Recent research shows that partner gender may affect how much effort we exert in teams, see Babcock et al. (2017). In my setting, using the insights from Experiment 1, I hypothesized that participants — in particular males — would lower their individual performance when working together with a female.

**Question 3 — Do women under-credit their contribution to group work?**

$$credit_{i,j} = contribution_{i,j} + Female\ Player_i + Female\ Partner_{i,j} + X_{i,j} + \epsilon_{i,j} \quad (3)$$

The outcome variable  $credit_{i,j}$  is defined for each individual  $i$  in each round  $j$  and can take on any value between 0 and 100 in increments of 0.5.  $Contribution_{i,j}$  ranges from 0 to 100 and is equal to the objective contribution that individual  $i$  made working on round  $j$ . The coefficient is the increase in credit claims when individual  $i$ 's contribution increases by 1 percentage point.  $Female\ Player_i$  is a dummy equal to one if subject  $i$  is a woman. The coefficient represents the percentage point difference in credit claims between females and males controlling for objective contribution and the set of controls in  $X$ . Based on the literature on gender differences in confidence, and Experiment 1, I expect this coefficient to be significant with a negative sign.

I then turn to the secondary test, where I analyze whether there is a positive sign on the partner’s gender, suggesting that individuals claim more credit when working with a woman than with a man. Based on e.g Gneezy, Niederle and Rustichini (2003) who shows that women like to compete with each other, I hypothesized that female under-claiming would be lower across all female teams. I examine whether this effect of the partner’s gender is smaller for women than for men by adding an interaction between gender and partner’s gender with a hypothesized negative sign. I use two-way clustering for the standard errors and cluster by participant and partner for all tests.

**Question 4 — Are men and women equally likely to reverse a partner’s move?  
Are they equally likely to have their moves reversed?**

$$reversion_{i,j,k} = Female\ Player_i + Female\ Partner_{i,j} + X_{i,j} + \epsilon_{i,j,k} \quad (4)$$

The outcome variable variable  $reversion_{i,j,k}$  is a dummy equal to one if move  $k$  made by individual  $i$  in round  $j$  is a reversion. I run this test on two separate data sets: first on the dataset containing all moves that might have been a correction, where the preceding move was bad, and then on all moves that might have been a bad reversion, where the preceding move was good, to measure gender differences in the propensity to make, and be the target of a correction and a bad reversion.

The coefficient on  $Female\ Player_i$  measures the percentage point difference in probability that a woman makes a correction, compared to a man. Based on the literature on female under confidence, I expect the coefficient on  $Female\ Player_i$  to be significant and negative, for both corrections and bad reversions. Moreover, based a recent paper showing that high quality female contributions to group work are rejected (see Terrell et al. (2017)) and insights from Experiment 1, I expect the coefficient on partner’s gender to be significant and negative, since I expect women to be less likely to be the target of a correction than men. I expect that the sign reverses when I instead consider bad reversions: women are expected to be more likely to be the target of bad reversions than men. I cluster by participant and partner.

## 5 Results

Table 1 presents summary statistics for the participants. Women and men are equally good at solving the puzzle on their own, are balanced on all background variables, and have similar priors on their future partners' ability before engaging in group work.

### 5.1 Team Success and Individual Performance.

#### Descriptives

In total, 679 puzzles were attempted and 505, or 74.4%, were solved. The average success rate is 66.3% in the first round and then increases in each round to 81.5% in the ninth and final round. The hardest board was solved only 43.1% of the time while the easiest was solved 92.9% of the time. Out of the 679 teams, 212 (31.3%) were male, 323 (47.6%) were mixed and the remaining 144 (21.2%) were female. The measure of individual performance is the number of good moves minus bad moves that a participant makes during a puzzle attempted with a partner. The individual performance ranges from -15 to 32. On average, participants have an individual performance of 4.8 (SD=4.2). The mean individual performance for men is 4.9 (SD= 4.3) and for women it is 4.7 (SD=4.1).

#### Does gender team composition matter for team success/individual performance? - Formal Test of Question 1 and 2

Table 2 shows that gender composition does not affect the success of a team: female, male and mixed teams are all equally successful in solving the puzzle. Turning to the question of whether partner gender affects individual performance, I show in the first column in table 3 that this is not the case. This table also shows that female individual performance within teams does not differ from male. Moreover, the second column of table 3 shows that gender composition does not affect individual performance. To summarize, women and men show equal levels of individual performance when working within a team regardless of the gender of their team mate. The gender

composition of a team does not determine team success. Thus, as we turn to credit claiming, we note that all tests of performance indicate equal ability and contribution by gender.

## 5.2 Credit Claims and Contributions to Successful Group Work.

### Descriptives

I focus on solved puzzles because shared accomplishments often play a part in important labor market decisions. Thus, I only use credit data from successful teams. This decision was made in advance and stated in the pre-analysis plan.

In total, 505 puzzles were solved. Since there are two participants in each team, there are 1,010 observations. Individual contributions to the solution vary from 0% to 100% and the mean contribution is 50.0% (SD=14.8). Figure 3 shows a histogram of individual contributions in solved puzzles, split by gender. Men on average contribute 50.0% (SD=16.0), and women on average contribute 50.0% (SD=13.3) to the solution. Thus, men and women contribute equally to the solution on average.

Turning to credit claims, these also range from 0% to 100%. The mean credit claimed for individual contributions to the solution is 54.7% (SD=17.6). For men the mean claim is 56.9% (SD=19.4), for women it's 52.0% (SD=14.7). Figure 4 shows a histogram of individual credit claims in solved puzzles, split by gender.

When considering all solved puzzles, the correlation between credit claimed and actual contribution is quite low, 0.18 for men and 0.07 for women. Figure 5 shows individual credit claims ranging from 0% to 100% on the y-axis plotted against individual contribution from 0% to 100% on the x-axis, split by gender. We see that both credit claims and contributions are spread, and that there is a weak positive correlation between claims and contributions indicating that participants understood and could apply the credit concept in solved puzzles to a certain extent. However, the correlation is stronger for men than women. Looking at figure 5, we also see that men consistently claim more credit than women for the same contribution, except for at very low levels of contribution.

I divide the data according to two pre-specified categories. First, I look at complex and simple solutions separately. A complex solution is one where the puzzle was solved in 1.7 times or more moves than the number of moves initially needed. For instance, if a puzzle initially needed 10 moves to be solved but a pair solved it in 17 moves or more, this solution is considered complex. If the same puzzle was solved in 16 moves or less, the solution is considered simple. This threshold was pre-registered and comes from the first data collection using this puzzle.

Figure 6 shows the credit claims and contributions in simple and complex solutions and their linear fit. Among women who worked on a complex solution, I find that the average claim is 50.1% (SD=17.6), and the average male claim is 57.9% (SD=19.7). In cases of simple solutions, the difference is smaller: the average female claim is 53.1% (SD=12.6) and male is 56.1% (SD= 19.2). The correlation between credit claims and contributions in complex solutions is 0.00 for women and 0.21 for men. In simple solutions, however, the correlation is 0.21 for women and 0.15 for men. In other words, in simple solutions women are approximately as likely as men to take credit for their contributions. In complex solutions however, there is no correlation between contributions and credit claimed for women. Men on the other hand, are approximately as likely to claim credit in both simple and complex team work. Thus, the extent to which women take credit for their contributions seems to be related to the complexity of the solution.

Second, I look at high contributors (those that contributed more than their partner), and low/equal contributors separately. Among high contributors, the average claim is 56.6 % (SD=17.7). Among women who contribute more than their partner, the average claim is 52.9% (SD= 14.6), and for male high contributors the average claim is 59.3% (SD= 19.3). If we consider those that contribute less than or equally to the solution, the male average claim is 55.4% (SD= 19.4) and the female average claim is 51.5 % (SD=14.7). Figure 7 shows average contribution claims in each quartile where Q1 corresponds to participants who contributed between 0 and 24 percent to the solution. As seen in this bar graph, in Q1, women and men claim the same amount of credit on average: 50%. The gender gap appears in Q2 and increases from there, and in Q4, the gap corresponds to 21 percentage points. Among female high contributors, the correlation between

contributions and claims is -0.01 and among male it is 0.22. In summary, the extent to which women take appropriate credit for their contributions to successful group work is related to their level of contribution, and the complexity by which the solution was reached.

Finally, in addition to the credit question, subjects were also asked after each puzzle whether or not they thought they would have done better on their own than with a partner. The answer alternatives ranged from *"Yes, I would have solved it much better on my own."* to *"No, I think I would have solved it much worse on my own."* Figure 8 shows the distribution of answers for women (black bars) and men (gray bars) separately. This figure shows that a larger share of women than men answer *"No, I would have done much/somewhat worse"*, and *"No, it would have stayed the same"*. While a larger share of men than women answer *"Yes, I would have done much/somewhat better"*. The average answer for women was *"No, I think it would have stayed the same."* whereas the average answer for men was *"Yes, I think I would have solved it somewhat better on my own."*. Note that I use all answers (solved and unsolved boards) here.

### **Do women claim less credit than men? - Formal test of Question 3**

To answer Question 3, I use data from solved boards (N=1,010). In table 4, I run a regression on credit claims, controlling for contribution levels and the standard set of controls. First note that the coefficient on contribution is statistically significant, meaning that participants updated their claims as their contribution levels increased. As seen in table 4, about 0.15 percent of an additional percent contribution is claimed by participants.

We see that, controlling for contributions, women claim 4.4 percentage points less credit than men ( $p=0.014$ ). I conclude that women under-credit their contributions. Table 4 also shows that partner gender does not have a significant effect on credit claims. The insignificant coefficient on the female team dummy in the second column shows that there are no interaction effects for partner gender: neither partner gender nor gender composition of the team matter for how much credit participants claim. Column 3 examines whether prior belief about future partners' contributions explains female under-crediting, by adding the belief that player  $i$  had about the

partner in round  $j$ 's future contribution before having worked together with that partner. Since the coefficient on female player remains negative and significant, I conclude that prior stereotypes, such as a belief that men are better at this game, are not driving the results.

In addition, in figure 9, I plot the distance between credit claims and contributions for men (dashed) and women (line) by round. The distance is larger for men in every round, and the gap doesn't close over time. In the appendix, I show that the main result that women under claim their contributions is robust when controlling for different characteristics of play, such as whether or not a correction was made during the game, see table 8. Also note in this table, that when we replace the contribution measure with our measure for individual performance — recall that this is the number of good minus bad moves made individually in each game — the result remains significant.

Next, I conduct two heterogeneity analyses by rerunning the same test, but for high and low/equal contributors, and simple and complex solutions separately. The results are seen in table 5. In column 1 and 2, I repeat the test using only data from high and low/equal contributors respectively. Comparing the coefficient on female player across these columns, we see that high-contributing women under-credit their contribution by 5.1 percentage points ( $p=0.012$ ), while the effect for low-performing women is weaker, 3.6 percentage points, and insignificant ( $p=0.083$ ). Column 3 and 4 repeat the same test but divides the data into complex and simple solutions. In column 3, we see that women who were part of a complex solution under-credit themselves by 5.5 percentage points ( $p=0.022$ ). In column 4, we see that the effect is directionally weaker for women who were part of a simple solution: 3.7 percentage points ( $p=0.047$ ). Thus, the finding that women under-credit their contributions is strongest among women who contribute more than their partner, and women who work towards complex solutions.

Finally, we turn to the responses of the question of whether the respondent would have been better on their own than with a partner. There were five answer alternatives, ranging from "*Yes, I would have solved it much better on my own*" — this answer was coded as 4 — to "*No, I think I would have solved it much worse on my own.*" — this answer was coded as 0. I run the same

specification as the one used for the tests of Question 3 on credit, but use answers for both solved and unsolved boards. The result is seen in table 6. Women report significantly lower values than men, suggesting that they are less likely to think they would have done better on their own than men.

## 5.3 Reversions

### Descriptives

I now focus on the moves of my 197 participants. In total, they made 20,190 moves for which a reversion was possible.<sup>18</sup> Out of these, about two thirds, or 12,991 moves are good and about one third, 7,199 moves, are bad. We will say that a correction is a reversion that undoes a bad previous move and a bad reversion is one that undoes a good previous move; there are 12,991 possibilities to make a bad reversion and 7,199 to make a correction. A reversion was made 1,330 times, or 6.5% of opportunities. Men made 849 corrections and women made only 481. Out of the 1,330 reversions, 737 were corrections and 593 were bad reversions. For corrections, men made 487 (11.3 % of the time that they have the chance) and women 250 (8.4 % of the time). For bad reversions, men made 362 (4.8%) and women made 231 (4.0%). Figure 10 shows the propensity to make a correction and bad reversion for women (black) and men (gray) separately. Turning to the question of whether men and women are equally likely to have their moves reversed by their partner, men were the target of a reversion 765 times (6.5 % of the time that they could have been), and women were the target 565 times (6.5%).

### Does gender matter for reversions? - Formal Test of Question 4

The test of Question 4 is presented in table 7. The coefficient on "female player" and "female partner" pertain to different components of Question 4. First, I conclude that women are 3.5

---

<sup>18</sup>Since a first move cannot be a reversion, all first moves have been removed from the dataset.



percentage points less likely than men to correct their partner ( $p=0.026$ ). Relative to the male baseline of making a correction at 11.3 %, this corresponds to a 31% gender gap in making corrections. Second, in column 2 I show that men and women are equally likely to make a bad reversion. Third, based on the coefficient on female partner dummy in column 1 and 2, I conclude that men and women are equally likely to be the target of a reversion. I conclude that men are more likely to correct a partner's mistake than women, but men and women are equally likely to have their moves reversed.

## 6 Discussion and Conclusion

This paper has analyzed gender differences in group work in three dimensions: how much credit individuals claim for a joint success, how performance is affected by the gender composition of a team, and how likely people are to correct their partner's mistakes. In order to answer these questions, I introduce a novel task which, to the best of my knowledge, is the first to clearly quantify individual contributions to shared work. I answer these questions using a laboratory experiment in which participants are randomized into pairs and solve a puzzle in teams of two. I show that despite the fact that women and men are equally good at solving this puzzle, both individually and as part of a team, women significantly and consistently under-credit their contribution to a shared success, with an average gap of 4.4 percentage points. I also find that women are less likely to correct a partner's mistake. Next, I discuss limitations, implications, and suggest future directions for this research.

While the puzzle allows for clean measurement of contribution, it does raise some potential concerns. First, the notion of a good move is limited in the sense that a good move is only preferable if you are playing with a partner who is also trying to solve the puzzle. Consider, for instance, playing with a partner who constantly moves to the right. The best move when working with such a partner is not necessarily a good move according to the algorithm. Instead you would adjust your moves to this erratic partner, possibly making effective moves that are labeled as bad by the algorithm. With such a partner, the player would be likely to feel that she made the larger

contribution, regardless of what the contribution measure says. This would only be a concern insofar as there is a gender difference in the propensity to be an erratic player and since there are no gender differences in any of the ability measures, this should not be the case.

Another limitation of using a good vs. bad move dichotomy is that some good moves are easier to make than others. Consider for instance the last move that solves the board, this move is clearly easier to spot than a good move made when you are far from the solution, yet both moves count equally. It is conceivable that the participant making good moves early in the game thinks that she deserves more credit than the participant making later good moves. Again, this would only be a concern to the extent that there are gender differences in making the first good move as opposed to the last good move. Given the structure of the game, where participants take turns, are randomized into making the first move, and the fact that there are no gender differences in ability in this game, this seems highly unlikely.

What are the implications of the finding that women under-credit their contributions? Under-crediting contributions to group work may have effects through at least two channels. First, if women attribute less credit than appropriate to themselves, this may lower their confidence, which could in turn have several negative consequences. For instance, they might not apply for promotions or negotiate higher wages if they undervalue their contributions to projects that went well. The second channel is that low self-attribution of credit might result in low external attributions of credit for shared work. In a group work setting where individual contributions are hard to monitor, an external evaluator might rely on credit claims when deciding who should be rewarded for successful shared work. In a team work environment where individual contributions are not immediately clear, it should be easier to get appropriate credit if you claim appropriate credit. At the same time, the benefits of being attributed credit are often substantial: tenure decisions, bonuses and promotions, and social recognition are often on the line when a boss makes decisions about who was responsible for successful team work.

Against this backdrop, the present results are cause for concern both for female advancement in the labor market and for organizations that might not be rewarding and promoting the right

people. Recent scholarship has found that women are given less than appropriate credit for their contributions to shared work, see Sarsons (2015). Even though it is not the main focus of this study, the indication that high-contributing women are especially likely to claim insufficient credit should be explored in future research. The fact that women in the highest quartile of contribution claim 21 percentage points less credit merits further attention. It is also important to note that the strength of this result is limited by the fact that the difference between high- and low/equal-contributing women is not significant. Studies with more power would be needed to establish whether high-contributing women claim significantly less credit than low/equal-contributing women. If this result were confirmed, it would have important implications, suggesting that the women who did the most work might not receive appropriate credit, in part perhaps because they do not make accurate claims about their contributions.

It is clear that behavior in a laboratory does not necessarily translate into behavior in the field. Nonetheless, there is reason to believe that the gender difference in claiming credit observed in this experiment represents a lower bound on the magnitude of the difference. In the field, incentives are often such that claiming more credit than appropriate is a winning strategy, in my experiment the monetary incentive is always aligned with telling the truth, since you earn more money the closer your credit claim is to your true contribution. Also, unlike the field, where claiming credit often involves speaking up, in the laboratory participants are sitting alone and are told that their credit claims will be invisible to their partner. Thus, the observed gender differences in speaking up in groups documented by Coffman (2014) are limited in this setting, increasing the likelihood that women will claim credit. However, because participants are told that their claims will be shown, together with their name and picture, to future participants in a follow-up study, I can not completely rule out that concerns about social desirability are driving the results. The implications of the finding that women are less likely to correct a partner's mistake are less clear. First, I think that it is important to note that due to their equal ability, women should be equally good at detecting mistakes. The lower propensity should not be driven by a lack in capacity. A correction affects team work in at least two ways. First, you move the team closer to the solution,

which is good. But second, because making a correction is tantamount to telling your partner that she was wrong, you might affect team dynamics, for instance, reducing motivation. Therefore it is not clear whether corrections are actually good for team work. To answer this question, a new study should be conducted in which the opportunity to correct is removed in a between-subjects treatment variation. This is an important topic, since corrections is a natural part of many working environments.

I plan to explore what drives my results in future work. One potential explanation that I will study is the fear of social backlash. It seems clear that a lot of research remains to be done to develop a better understanding of how, why, and when we succeed in working well together and appreciate our own contributions and the contributions of others. Answering these questions is critical, given the significance of group work in modern work and the continued gender inequality in the labor market.

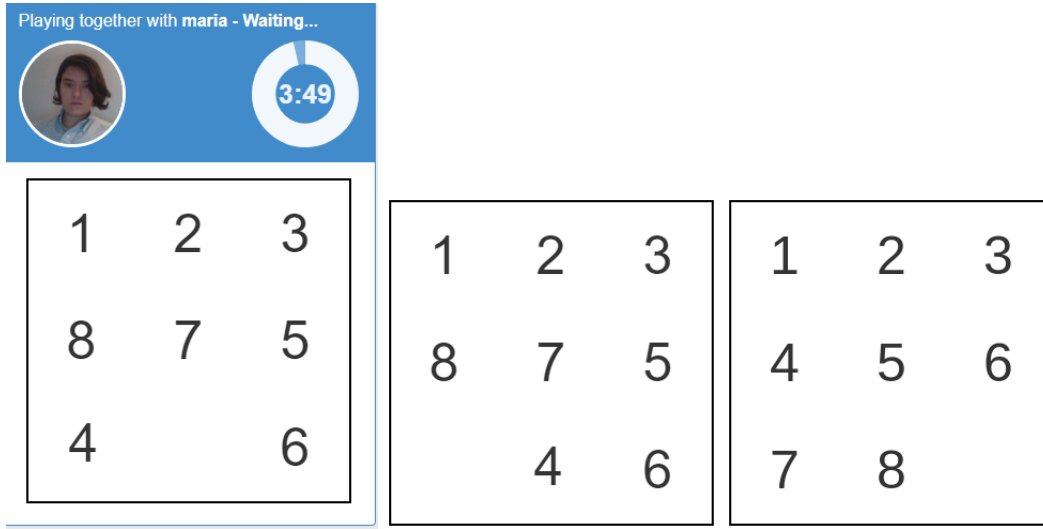
## References

- Anderson, Cameron, Sebastien Brion, Don A Moore, and Jessica A Kennedy.** 2012. “A status-enhancement account of overconfidence.” Journal of personality and social psychology, 103(4): 718.
- Archer, Aaron F.** 1999. “A modern treatment of the 15 puzzle.” The American mathematical monthly, 106(9): 793–799.
- Azmat, Ghazala, and Barbara Petrongolo.** 2014. “Gender and the labor market: What have we learned from field and lab experiments?” Labour Economics, 30: 32–40.
- Babcock, Linda, Maria 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): 714–47.
- Blau, Francine D, and Lawrence M Kahn.** 2017. “The gender wage gap: Extent, trends, and explanations.” Journal of Economic Literature, 55(3): 789–865.
- Born, Andreas, Eva Ranehill, and Anna Sandberg.** 2018. “A Man’s World? The Impact of a Male Dominated Environment on Female Leadership.”
- Bundy, Alan, and Lincoln Wallen.** 1984. “Breadth-first search.” In Catalogue of Artificial Intelligence Tools. 13–13. Springer.
- 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): 448–74.
- Gneezy, Uri, Muriel Niederle, and Aldo Rustichini.** 2003. “Performance in competitive environments: Gender differences.” The Quarterly Journal of Economics, 118(3): 1049–1074.

- Haynes, Michelle C, and Madeline E Heilman.** 2013. “It had to be you (not me)! Womens attributional rationalization of their contribution to successful joint work outcomes.” Personality and Social Psychology Bulletin, 39(7): 956–969.
- Hoogendoorn, Sander, Hessel Oosterbeek, and Mirjam Van Praag.** 2013. “The impact of gender diversity on the performance of business teams: Evidence from a field experiment.” Management Science, 59(7): 1514–1528.
- Hossain, Tanjim, and Ryo Okui.** 2013. “The binarized scoring rule.” Review of Economic Studies, 80(3): 984–1001.
- Ivanova-Stenzel, Radosveta, and Dorothea Kübler.** 2011. “Gender differences in team work and team competition.” Journal of Economic Psychology, 32(5): 797–808.
- 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.
- Niederle, Muriel, and Lise Vesterlund.** 2011. “Gender and competition.” Annu. Rev. Econ., 3(1): 601–630.
- Sarsons, Heather.** 2015. “Gender differences in recognition for group work.” Harvard University, 3.
- Terrell, Josh, Andrew Kofink, Justin Middleton, Clarissa Raine, Emerson Murphy-Hill, Chris Parnin, and Jon Stallings.** 2017. “Gender differences and bias in open source: Pull request acceptance of women versus men.” PeerJ Computer Science, 3: e111.

## A Figures

Figure 1: The puzzle



*Notes:* Panel a of this figure shows a scenario in which a player is waiting for her partner Maria to make a move. In this example scenario, at least seven steps remain to solve the board. Maria has two options: she can either move 7 down (by clicking on the tile that says 7) or 4 to the right. Each move is either good in the sense that it moves the team closer to the solution, or bad in the sense that it moves the team further away from the solution. In this case the first option (7 down) is good, and the second option (4 right) is bad. So if Maria chooses to move 7 down, at least six moves remain to solve the puzzle, whereas if she moves 4 right at least 8 moves are needed to reach the solution. Panel b shows the configuration if Maria moves 4 right in the scenario seen in panel a. If a participant thinks that her partner made a mistake, she is free to reverse her partner's move. Panel c shows the configuration when the board is solved. The goal is to get all tiles in numerical order, leaving the bottom right corner empty. This is achieved by taking turns in moving tiles within each team until the goal state is reached.



Figure 2: The credit question

Sliding Puzzle result for carl

You solved the sliding puzzle R-A1 with MARIA.  
Please make an estimation ranging from 0% to 100% of what you think your contribution in solving sliding puzzle R-A1 was.

Show more info

Your contribution:

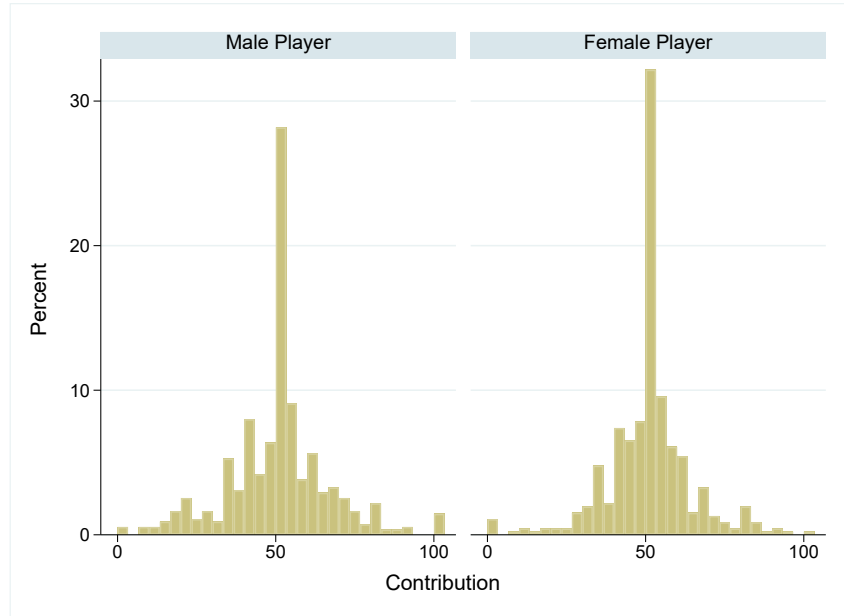
(Not visible to your partner)

Your contribution

Do you think you would be better on your own?

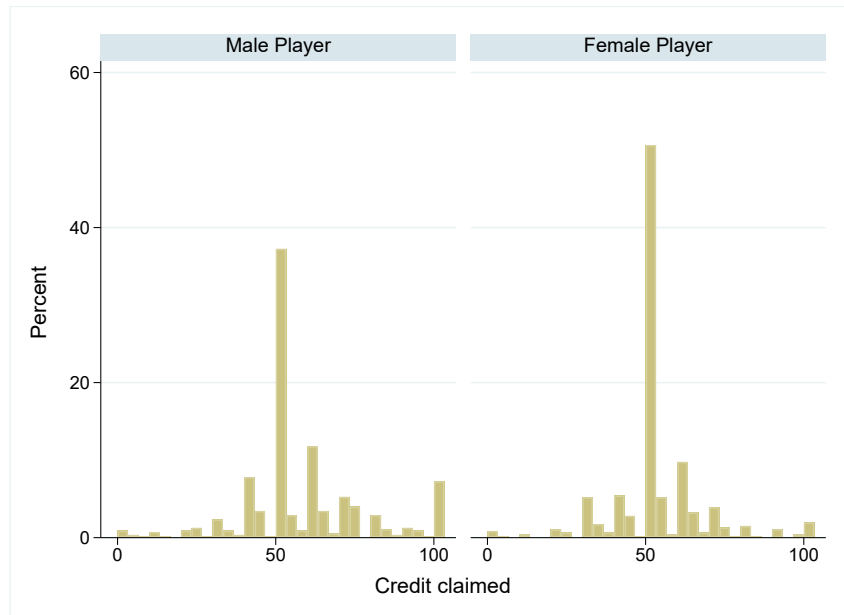
*Notes:* The figure shows an example credit question for a fictive participant named Carl who solved a puzzle together with Maria. Participants answer from 0% to 100% how much they think they contributed to the solution. The correct answer in this case, is the sum of good minus bad moves that a Carl made in this game relative to Maria. The question is incentivized using the binarized scoring rule of Hossain and Okui (2013). Participants earn more money the closer their credit claim is to their true contribution. Below the credit question is a subjective question without incentives asking whether the participant thinks she would have done better on her own, there are five answer alternatives ranging from "Yes, I would have solved it much better on my own." to "No, I think I would have solved it much worse on my own."

Figure 3: Histogram of individual contributions in solved puzzles



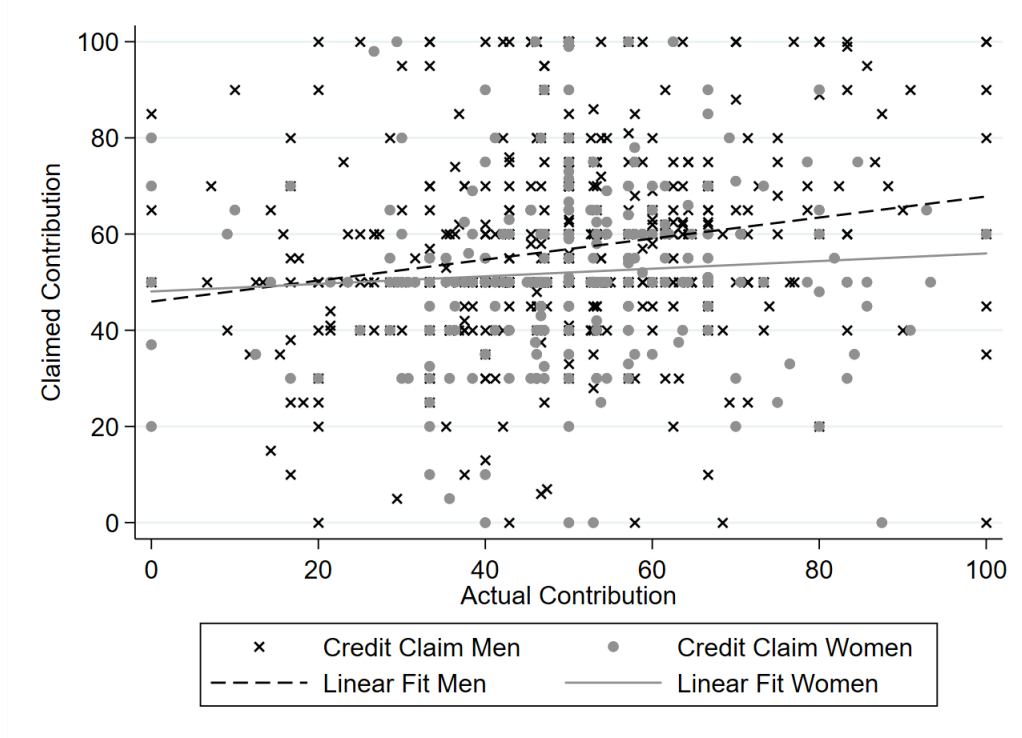
*Notes:* The figure shows the histogram of individual contributions in solved puzzles, split by gender. Individual contributions range from 0% to 100%. The mean contribution in a solved puzzle is 50.0% (SD=14.8). Men on average contribute 50% (SD=16.0), and women on average contribute 50.0% (SD=13.3) to the solution. Thus, men and women contribute equally to the solution on average.

Figure 4: Histogram of individual credit claims in solved puzzles



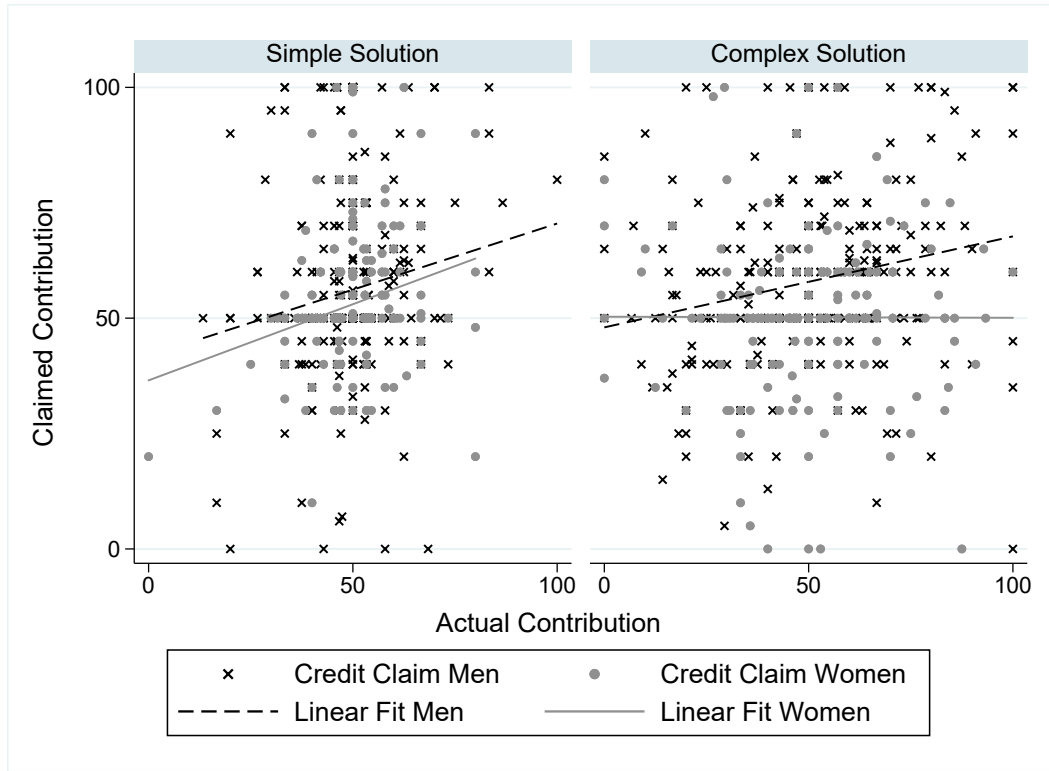
*Notes:* The figure shows the histogram of individual credit claims in solved puzzles, split by gender. Individual credit claims range from 0% to 100%. The mean credit claim in a solved puzzle is 54.7% (SD=17.6). For men, the mean claim is 56.9% (SD=19.4), for women it's 52.0% (SD=14.7).

Figure 5: Gender differences in claiming credit



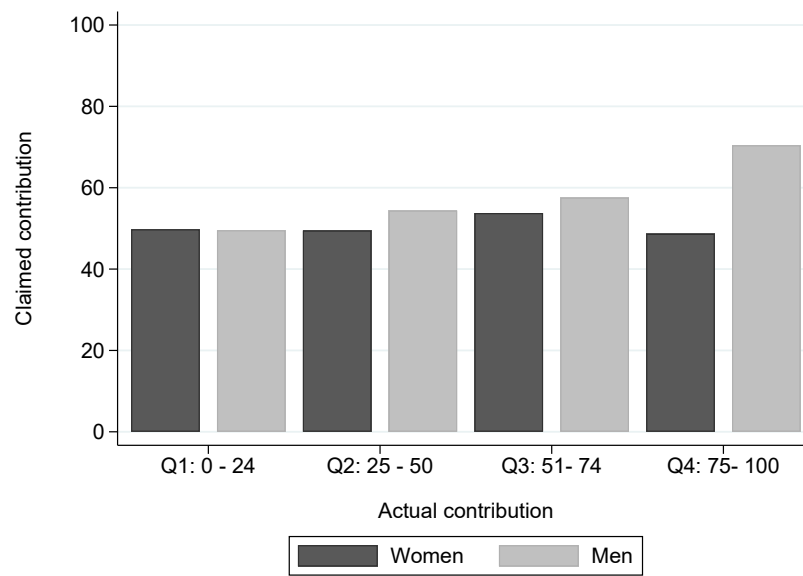
*Notes:* This figure shows gender differences in claiming credit. We see individual credit claims ranging from 0% to 100% on the y-axis, plotted against individual contribution from 0% to 100% on the x-axis, for men (x) and women (dot). For example, we see that among participants who contribute 0% to the solution, the lowest claimed contribution is 20% and made by a woman, and the highest claim is 85% and made by a man. As seen in the figure, both women and men have a weak positive correlation between credit claims and contributions in solved puzzles. For women, this correlation is 0.07 and for men it is 0.18. Except for at very low levels of contribution, women consistently claim less credit for the same contribution than men.

Figure 6: Gender differences in claiming credit in simple and complex solutions



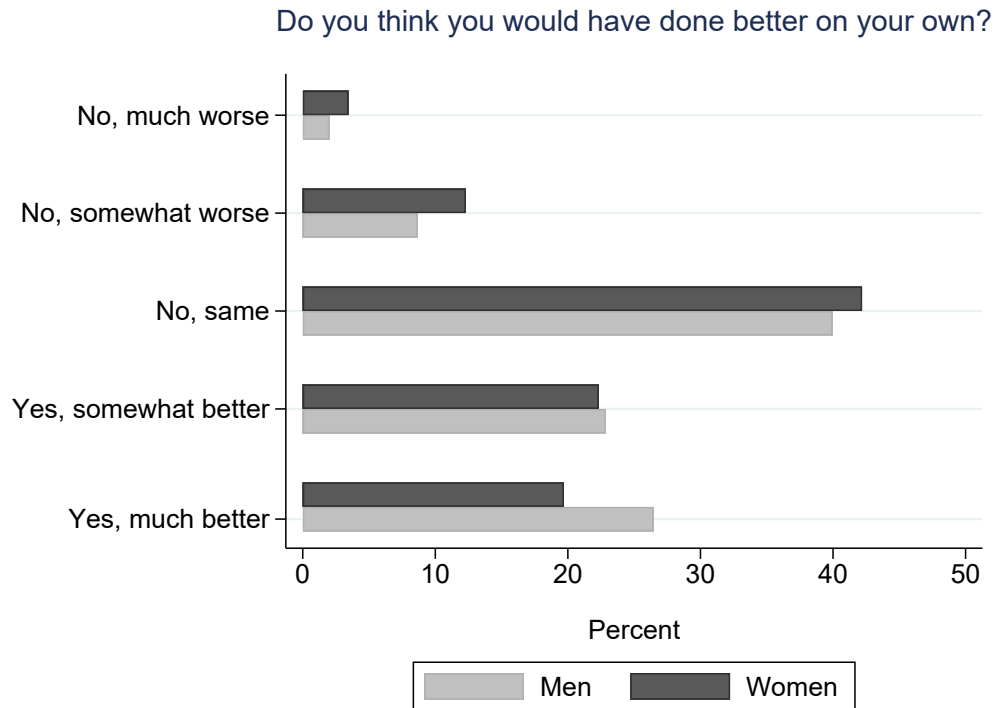
*Notes:* This figure shows gender differences in claiming credit in complex and simple solutions. We see individual credit claims and contributions in solved puzzles for men (x) and women (dot) in simple (left) and complex (right) solutions respectively. A complex solution is one where the solution was reached in more than or equal to 1.7 times the number of steps initially needed. For instance, if a puzzle initially required at least 10 moves to be solved, and it was solved in 18 moves it is considered a complex solution. If instead it was reached in 11 steps, it is considered a simple solution. As seen in the left panel, there is a positive correlation between credit claims and contribution in simple solutions. For women who contributed to a simple solution, this correlation is 0.21 and for men it is 0.15. In the complex solutions however, only men seem to increase their claims with their contributions, as seen in the right panel. The correlation for men in complex solutions is 0.21 whereas for women there is no positive correlation ( $\text{corr}=0.00$ ). The threshold of 1.7 comes from a previous study using this puzzle and was pre-registered.

Figure 7: Average claims in each quartile of contribution



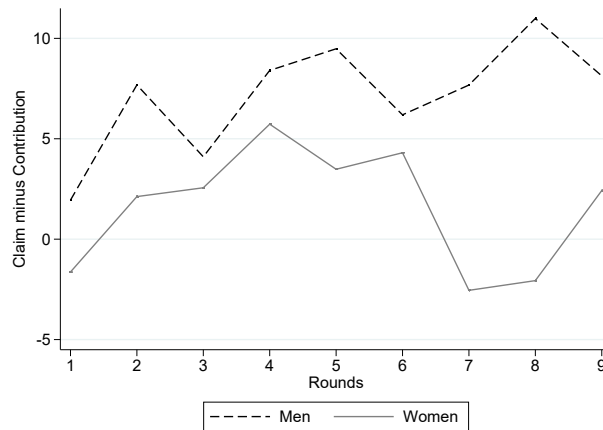
*Notes:* The figure shows average contribution claims in each quartile where Q1 corresponds to participants who contributed between 0 and 24 percent to the solution. As seen in this bar graph, in Q1, women and men claim the same amount of credit on average: 50%. The gender gap appears in Q2 and increases from there, and in Q4, the gap corresponds to 21 percentage points.

Figure 8: Subjective sentiment about contribution



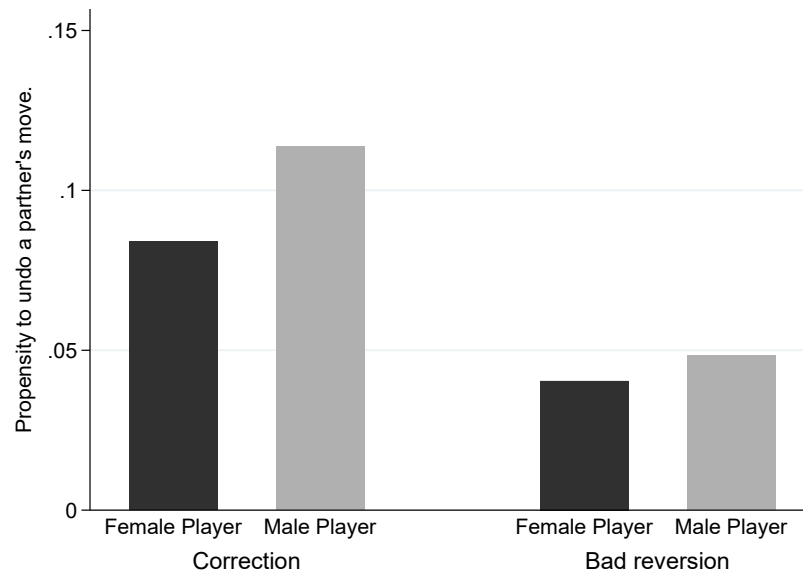
Notes: The figure shows the distribution of female and male answers to the question "Do you think you would be better on your own?". There were five answer alternatives, ranging from "Yes, I would have solved it much better on my own" to "No, I think I would have solved it much worse on my own." This figure shows that a larger share of women than men answer "No, I think I would have solved it much/somewhat worse" and "No, it would have stayed the same.". While a larger share of men than women answer "Yes, I would have solved it much/somewhat better".

Figure 9: Distance between claims and contributions by round



Notes: The figure shows the distance between credit claims and contributions for men (dashed) and women (line) by round. The distance is larger for men in every round, and the gap doesn't close over time.

Figure 10: Propensity to undo a partner's move



*Notes:* This figure shows the propensity of participants to undo a partner's move. A correction is a reversion that undoes a bad preceding move and a bad reversion is one that undoes a good preceding move. To the left, we see the propensity to make a correction when it's possible (the preceding move was bad) for a female player (black) and a male player (gray) separately. As seen in this figure, men have a higher propensity to correct a partner's mistake. To the right, we see the propensity to reverse a partner's good move. There are no gender differences in making a bad reversion.

## B Tables



Table 1: Summary statistics

		Men	Women	p-value [H <sub>0</sub> : M=W]
Socio-demographic characteristics				
Level of Education	High School	14.7%	17.0%	0.50
	Bachelor Degree	55.0%	45.5%	
	Masters Degree	27.5%	31.8%	
	Other	2.8%	5.7%	
Geographic background	Africa	1.8%	2.3%	0.30
	Asia	31.2%	21.6%	
	Europe	1.8%	5.7%	
	North America	64.2%	70.5%	
	South America	0.9%	0.0%	
Ethnicity	White	40.4%	48.9%	0.14
	African American	11.0%	5.7%	
	Native American	1.8%	0.0%	
	Asian	33.9%	27.3%	
	Latino	8.3%	5.7%	
	Mixed Race	3.7%	6.8%	
	Other	0.9%	5.7%	
Age	Mean	32.90	30.89	0.28
	(S.D)	(13.68)	(11.82)	
Field of study	Economics	4.6%	6.8%	0.10
	Political Science	10.1%	8.0%	
	Mathematics	9.2%	5.7%	
	Psychology	3.7%	11.4%	
	Humanities	9.2%	11.4%	
	Other Social Sciences	6.4%	13.6%	
	Other Natural Sciences	13.8%	15.9%	
	Other	43.1%	27.3%	
Pre-group work				
# of practice puzzles solved individually	Mean	6.88	7.42	0.17
	(S.D)	(2.84)	(2.63 )	
Prior beliefs on future partner's average contribution	Mean	49.41	50.00	0.83
	(S.D)	(21.29)	(16.55)	
N		109	88	

*Notes:* This table presents summary statistics for the participants. Women and men are equally good at solving the puzzle on their own, are balanced on all background variables, and have similar priors on their future partners' ability before engaging in group work.

Table 2: Team gender composition and team success

Probability of solving the puzzle	
Female Team	-0.0417 (0.0459)
Mixed Team	-0.0312 (0.0368)
Constant	0.574** (0.188)
Controls	Yes
Observations	1357
$R^2$	0.174
Standard errors in parentheses	
* $p < 0.05$ , ** $p < 0.01$ , *** $p < 0.001$	

*Notes:* This table shows a regression on the individual data set ( $N = 1,357$ ) from all nine rounds. I use male teams as a baseline and measure the relative success of mixed and female teams. We see that gender composition does not affect the success of a team. The regression is clustered on the team and player level.

Table 3: Team gender composition and individual performance

Individual performance within teams		
Female player	-0.300 (0.227)	-0.254 (0.309)
Female partner	0.201 (0.240)	0.247 (0.325)
Female team		-0.0998 (0.417)
Constant	5.809*** (1.110)	5.791*** (1.121)
Controls	Yes	Yes
Observations	1357	1357
$R^2$	0.126	0.126
Standard errors in parentheses		
* $p < 0.05$ , ** $p < 0.01$ , *** $p < 0.001$		

*Notes:* The outcome variable in this table is the sum of good net bad moves that an individual makes within a team. This table shows that partner gender does not affect individual performance. We also see that female individual performance within teams does not differ from male. Moreover, the second column of table shows that gender composition does not affect individual performance. The regression is clustered on the player and the partner.

Table 4: Gender differences in claiming credit for a joint success

	Credit claimed		
Contribution	0.151** (0.0468)	0.152** (0.0467)	0.153** (0.0472)
Female player	-4.406* (1.794)	-4.916** (1.859)	-4.381* (1.791)
Female partner	0.349 (1.145)	-0.158 (1.677)	0.403 (1.148)
Female team		1.104 (2.307)	
Prior			0.0213 (0.0223)
Constant	58.62*** (14.22)	58.66*** (14.22)	57.36*** (14.18)
Controls	Yes	Yes	Yes
Observations	1010	1010	1006
$R^2$	0.147	0.147	0.149

Standard errors in parentheses

\*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$

*Notes:* The outcome variable in this table is credit claimed in a solved puzzle, which ranges from 0% to 100%. This table shows that women under credit their contribution with 4.4 percentage points compared to men. The coefficient on "Female Partner Dummy" in the first column tests the secondary hypothesis on the effect of partner gender. Since this coefficient is insignificant, we conclude that partner gender does not affect credit claims. Column 2 tests for gender differences in the effect of partner gender by adding the "Female Team Dummy" which is equal to one if both players in a team are female. We see that there are no gender interaction effects. Column 3 repeats the test in column 1 and adds the prior belief of how much participants think their future partners will contribute. We conclude that priors do not explain why women under credit their contributions. All regressions are clustered by player and partner.

Table 5: Heterogeneity analyses

	Credit claimed			
	High	Low/Equal	Complex	Simple
Contribution	0.173 (0.107)	0.131 (0.0678)	0.107* (0.0479)	0.288*** (0.0840)
Female player	-5.115* (2.030)	-3.598 (2.076)	-5.515* (2.415)	-3.695* (1.860)
Female partner	1.516 (1.802)	-0.459 (1.456)	0.182 (2.176)	-0.699 (1.468)
Constant	56.79** (17.53)	61.07*** (13.47)	56.27*** (15.65)	51.62*** (15.22)
Controls	Yes	Yes	Yes	Yes
Observations	368	642	402	608
$R^2$	0.233	0.150	0.214	0.163

Standard errors in parentheses

\*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ 

*Notes:* In column 1 and 2 we repeat the test on gender differences in claiming credit for a joint success using data only from high (those that contribute more than their partner to the solution) and low contributors respectively. Comparing the coefficient on "Female Player Dummy" in column 1 and 2, we note that high contributing women under credit their contribution with 5.1 percentage points ( $p=0.012$ ) whereas the effect for low contributing women is weaker, 3.6 percentage points, and insignificant ( $p=0.083$ ). Column 3 and 4 repeat the same exercise but divides the data into complex solutions and simple solutions. In column 3, we see that women who were part of a complex solution under credit themselves with 5.5 percentage points ( $p=0.022$ ). In column 4, we see that the effect is weaker for women who were part of a simple solution, 3.7 percentage points ( $p=0.047$ ).

Table 6: Gender differences in subjective beliefs

Confidence in own's performance if playing alone (0 = high, 4 = low)	
Female player	-0.245* (0.101)
Female partner	-0.0268 (0.0558)
Constant	1.891*** (0.363)
Controls	Yes
Observations	1357
$R^2$	0.128
Standard errors in parentheses	
* $p < 0.05$ , ** $p < 0.01$ , *** $p < 0.001$	

*Notes:* This table shows a regression using data on the answer to the question "Do you think you would be better on your own?" in both solved and unsolved puzzles ( $N = 1,357$ ). There were five answer alternatives, ranging from "Yes, I would have solved it much better on my own." — this answer was coded as 4 — to "No, I think I would have solved it much worse on my own." — this answer was coded as 0. Women report significantly lower values than men, suggesting that they are less likely to think they would have done better on their own. The regression is clustered by player and partner.

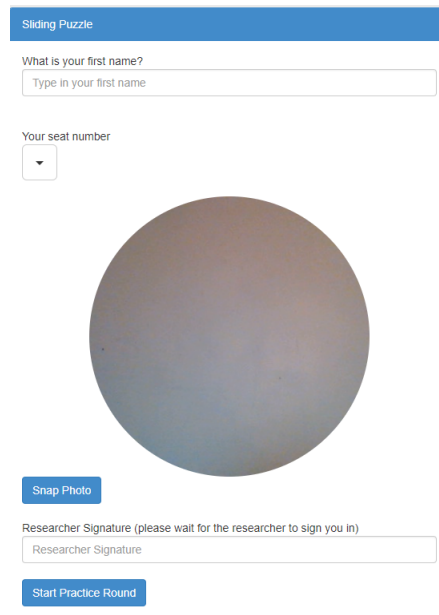
Table 7: Gender differences in the propensity to reverse a partner's move

Propensity to reverse a partner's move		
	Correction	Bad reversion
Female player	-0.0351* (0.0158)	-0.0122 (0.00843)
Female partner	0.0124 (0.0131)	-0.00341 (0.00489)
Constant	0.140* (0.0658)	0.0681** (0.0262)
Controls	Yes	Yes
Observations	7199	12991
Standard errors in parentheses		
* $p < 0.05$ , ** $p < 0.01$ , *** $p < 0.001$		

This table analyzes the propensity of participants to undo a partner's move. A correction is a reversion that undoes a bad preceding move and a bad reversion is one that undoes a good preceding move. We see that female players are 3.5 percentage points less likely to make a correction than males, this corresponds to a 31% gender gap in correcting a partner's mistake. There are no gender differences in bad reversions and there and men and women are equally likely to be the target of a reversion.

## C Appendix figures

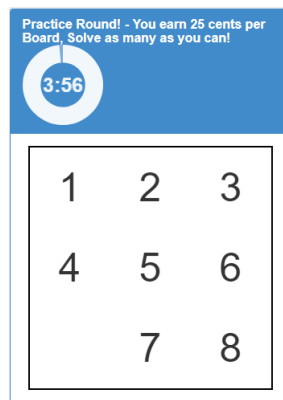
Figure 11: The sign in screen



The sign-in screen has a blue header with the text "Sliding Puzzle". Below the header, there is a text input field labeled "What is your first name?" with a placeholder "Type in your first name". Below this is a dropdown menu labeled "Your seat number". In the center of the screen is a large, circular, blurry image of a person. Below the image is a blue button labeled "Snap Photo". Below the button is a text input field labeled "Researcher Signature (please wait for the researcher to sign you in)" with a placeholder "Researcher Signature". At the bottom is a blue button labeled "Start Practice Round".

*Notes:* The figure shows the sign in screen. Here, participants enter their first name, seat number and take a picture of themselves. The researcher goes around and checks that all information is correct. Once all participants have entered correct information, the researcher signs in participants by entering the researcher signature and the practice round begins.

Figure 12: The first practice board

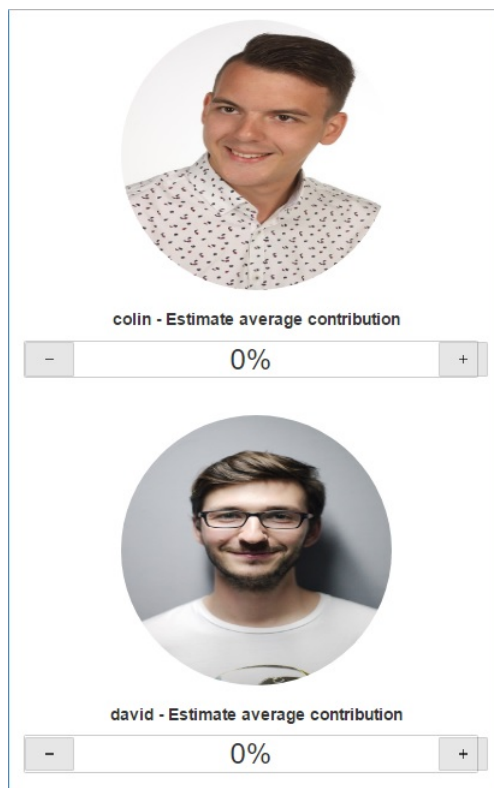


The practice board has a blue header with the text "Practice Round! - You earn 25 cents per Board. Solve as many as you can!". Below the header is a circular timer showing "3:56". Below the timer is a 3x3 grid of numbers. The numbers are arranged as follows:

1	2	3
4	5	6
	7	8

*Notes:* The figure shows the first and easiest practice puzzle. In the practice round, participants have 4 minutes to solve as many practice puzzles as possible. They earn 25 cents per solved practice puzzle, and are urged to solve as many as they can. Practice puzzles start very easy, and progress to become very hard. In total there are 15 practice puzzles. The number of practice puzzles solved in 4 minutes is my measure of individual ability.

Figure 13: The priors question



The image shows a digital interface for a 'priors question'. It contains two sections, one for 'colin' and one for 'david'. Each section features a circular profile picture of a man. Below each picture is a text label: 'colin - Estimate average contribution' and 'david - Estimate average contribution'. Under each label is a horizontal slider bar with a minus sign on the left and a plus sign on the right. The slider for Colin is set to '0%', and the slider for David is also set to '0%'.

*Notes:* The figure shows an example priors question. In the figure, we see how a player is asked to estimate the average contribution of the two fictive future partners Colin and David, ranging from 0% to 100%. The player is asked to do this before having worked together with Colin and David, and can thus only base her estimate on the picture and first name.



## D Appendix tables

Table 8: Mechanisms: Game Control. The outcome variable is credit claimed in a solved puzzle (N=1,010).

Credit claimed							
Performance	1.095*** (0.308)	1.095*** (0.307)	1.022*** (0.300)	1.018*** (0.302)	1.024*** (0.300)	1.012*** (0.298)	0.980** (0.327)
Female player	-4.380* (1.789)	-4.383* (1.811)	-4.342* (1.810)	-4.346* (1.810)	-4.424* (1.816)	-4.399* (1.795)	-4.319* (1.791)
Female partner	0.283 (1.147)	0.281 (1.140)	0.267 (1.144)	0.259 (1.139)	0.186 (1.130)	0.201 (1.132)	0.240 (1.130)
Correction		-0.0256 (1.212)	-1.363 (1.591)	-1.221 (1.656)	-1.687 (1.619)	-0.411 (1.641)	-0.495 (1.690)
Nr. of corrections			0.568 (0.489)	0.618 (0.533)	0.916 (0.554)	0.836 (0.550)	0.761 (0.668)
Loop				-0.520 (1.765)	1.168 (1.982)	1.772 (2.016)	1.734 (2.070)
# of loops					-0.432 (0.259)	-0.186 (0.259)	-0.248 (0.288)
time						-0.0412** (0.0142)	-0.0474** (0.0166)
# moves							0.0441 (0.0671)
Constant	56.78*** (14.36)	56.79*** (14.39)	57.62*** (14.34)	57.70*** (14.36)	58.44*** (14.26)	62.19*** (14.31)	62.51*** (14.33)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1010	1010	1010	1010	1010	1010	1008
$R^2$	0.147	0.147	0.149	0.149	0.152	0.162	0.157

Standard errors in parentheses

\*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$

*Notes:* In this table, we have added the game controls sequentially to check whether they could be driving female under crediting. As seen in the table, the coefficient on "Female Player" is significant and negative and similar in size in all specifications. We conclude that the finding that women under credit their contributions is robust to, and can not be explained by any of the game controls. All regressions are clustered at player and partner.