Gender stereotypes and homophily in team formation

Antonio Cabrales, Lorenzo Ductor, Ericka Rascón-Ramírez, and Ismael Rodriguez-Lara[*]

February 20, 2025

Abstract

Women often find themselves in teams that hinder their productivity and earnings. We analyze the role of homophily and gender stereotypes in preferences for team formation and examine the effect of information on changing these preferences. We find that women are expected to perform better in female-type tasks (such as text and emotion-recognition). However, people prefer forming teams with their same gender. Our findings suggest that information can mitigate -but it does not eliminate- the influence of homophily on team formation.

Keywords: gender differences, expectations, collaboration, network formation, team production.

JEL Class.: C91, D03, D60, D81.

[*] Cabrales: Department of Economics, Universidad Carlos III Madrid, Spain (email: antonio.cabrales@uc3m.es); Ductor: Department of Economics Theory and History, Universidad de Granada (lductor@gmail.com); Spain; Rascón-Ramírez: Department of Economics, Centro de Investigación y Docencia Económicas (CIDE), Mexico; and Department of Accounting, Finance & Economics, Middlesex University London (E.Ramirez@mdx.ac.uk); Rodriguez-Lara: Department of Economics, Universidad de Malaga, Spain; and Economic Science Institute (ESI), Chapman University, USA (ismaelrl@uma.es). We are grateful to Gary Charness, participants of the 2022 ESA World Meeting Boston, 2022 ESA European Meeting Bologna, 2023 and 2024 Asia-Pacific ESA Conference, 2024 SEET Meeting Malaga, 2024 ESA World Meeting Bogota, 2024 ESA European Meeting Helsinki, and seminar audience of University of Malaga, University of Verona, University California Santa Barbara, Centro de Investigación y Docencia Económicas and National University of Singapore for their valuable comments. We acknowledge the financial support of the *Fundación Ramón Areces*, and the Ministry of Science and Innovation of Spain, Grants PID2022-142943NB-I00 and and PID2021-126892NB-I00, funded by MICIU/AEI/ 10.13039/501100011033 and, ERDF/EU, as well as the Agencia Estatal de Investigación under María de Maeztu Programme CEX2021-001181-M.

1 Introduction

In the last decade, a large body of research has identified factors that contribute to the stubborn persistence of gender gaps.[1] In recent years, there is a strand in the literature that points out that women may work in teams that are detrimental for their productivity and earnings (Lindenlaub and Prummer, 2020; Ductor et al., 2023; Ductor and Prummer, 2024). In academia, Ductor et al. (2023) show that women have fewer collaborators and collaborate more with the same individuals, which can explain 18% of the productivity gap in research output in economics and all the research output gap in sociology.[2]

This paper deepens our understanding of the team element of the gender labor market gap by examining the role of team preferences and information. We hypothesize that men and women exhibit different preferences for team members and that these preferences depend on the type of task. We also hypothesize that by providing information the preferences may be modified. Our hypotheses are based on existing evidence on discrimination and information. For example, Reuben et al. (2014) show that women are discriminated in tasks for which men are perceived to perform better. Moreover, this bias persists to some extent when correct information about performance is provided.

We conducted three (preregistered) experiments[3] with more than 2,800 participants recruited from Prolific (Palan and Schitter, 2018). In the experiments, we measured performance, beliefs, and team formation in a series of gender-stereotyped tasks.[4] Our female stereotypical tasks involve completing paragraphs by selecting the most appropriate word (Text) and identifying emotions from facial expressions (Emotion). Our male stereotypical tasks involve solving numerical exercises (Math) and matching or distinguishing between

[1] See Blau and Kahn (2000), Goldin et al. (2006), Azmat and Petrongolo (2014) and Lalanne and Seabright (2016). Lalanne and Seabright (2016) find that women account for only 2.4% of CEO in US firms. In 2022, women who worked full-time in US or the EU had a median weekly earnings of 82% of men's median weekly earnings.

[2] See, among others, Bagues and Esteve-Volart (2010), Card et al. (2020), Fenoll and Zaccagni (2022) or Ductor and Prummer (2024) for evidence that the gender composition of teams affects productivity.

[3] Experimental techniques have already proven to be useful for understanding behavior in the workplace (Charness and Kuhn, 2010; Herbst and Mas, 2015; Azmat and Petrongolo, 2014; Averett et al., 2018)

[4] The hypotheses, experimental design, sampling, preanalysis and exclusion criteria for our experiments can be consulted at AsPredicted.org (#60390 and #153944). Ethical approval was obtained from the Universidad de Granada (2060/ CEIH/2021) and the Universidad de Malaga (CEUMA 45-2023-H).

3D figures (Rotation). We choose these tasks because women are often associated with verbal and emotional intelligence skills (Helgeson (2020); Brody (1997)), while men are often associated with spatial and numerical skills (Correll, 2001; Rudman and Kilianski, 2001; Kiefer and Sekaquaptewa, 2007; Bordalo et al., 2019; Vandenberg and Kuse, 1978; Maeda and Yoon, 2013). However, empirical evidence suggests that these stereotypes do not necessarily reflect actual performance; e.g., in verbal or spatial tasks (Hyde et al., 1990; Niederle and Vesterlund, 2007; Niederle et al., 2013). Our approach therefore relies on tasks where beliefs about performance may be inaccurate.[5] This, in turn, allows us to assess the effect of providing correct information on team composition.

Our three experiments proceed as follows. In our *task experiment*, men and women work individually in one of the gender-stereotyped tasks (Tex, Emotion, Math or Rotation) for 15 minutes. We measure expectations in our *belief experiment* by asking external observers to predict the performance of men and women in the *task experiment.* Our *team experiment* allows men and women to express their preferences for team formation as participants must form a team by selecting the avatars of men and women who participated in the task experiment.

One of our main objectives is to establish how the decision to work with others is affected by preferences (taste-based discrimination) and beliefs (statistical discrimination) about the productivity of others. For this reason, our team experiment consists of two different treatments. In one treatment, participants select team members without any information about the performance of others, allowing both preferences and beliefs to influence their decisions. We can assess the role of beliefs in team formation in this case by relying on our belief experiment. In the second treatment, participants are provided with performance information from the task experiment. Since this relieves the team formation decision from the effect of beliefs, we can assess more clearly the impact of preferences.

Our design choices allow our analysis to focus on our factors of interest, preferences, and information. They eliminate others that may affect the formation of teams in real-world settings. In our study, participants do not have different opportunities to form links. They also do not have any cost associated with the creation of such links. Because our experiment

[5] Our preregistered hypotheses posit that women will excel in the Emotion task (Lambrecht et al., 2014; Kessels et al., 2014), while men will excel in the Math task (Hyde et al., 2008; Hyde and Mertz, 2009; Iriberri and Rey-Biel, 2019; Dossi et al., 2021).

does not involve interaction with other participants, soft skills or personality traits should not play any role in the selection of team members. In this respect, there is no role for prosocial or free-riding behavior in our team experiment. This is important as women are often more prosocial and cooperative in teams (Babcock et al., 2022; Charness and Rustichini, 2011) and react differently to peer pressure (Beugnot et al., 2019) than men.

Our main result is that homophily plays an important role in team formation: men prefer to form teams with a majority of men, and women prefer to form teams with a majority of women.[6] This occurs regardless of the task at stake. In addition, the tendency to select team members of the same gender seems to be based on taste (preferences), as it cannot be explained by beliefs about performance. Our results also suggest that information can mitigate (but does not eliminate) the influence of homophily on team formation.

In terms of expectations, we find that women are expected to perform better in female-stereotypical tasks (Text or Emotion), but there are no expected gender differences in average performance in Math and Rotation tasks. With regard to the actual average performance observed in the *task experiment*, we find no gender differences except for the Emotion task, in which women perform better. However, we also observe that men are more frequently found at the extremes of the performance distribution for Text and Math, indicating that men are riskier choices for these tasks.

1.1 Literature

Our work connects to several strands of the literature.

Our work contributes to a better understanding of gender inequality in the workplace (Bertrand, 2011; Blau and Kahn, 2016; Averett et al., 2018). Gender differences in labor market outcomes can be attributed to three main factors (Azmat and Petrongolo, 2014): (1) differences in characteristics, preferences, skills, or team composition leading to lower productivity (Eckel and Grossman, 2008; Croson and Gneezy, 2009; Lindenlaub and Prummer, 2020; Ductor et al., 2018; Ductor and Prummer, 2024); (2) greater family constraints faced by women (Bertrand et al., 2010; Albanesi and Olivetti, 2009; Adda et al., 2011); or (3) discrimination against women in that women with the same preferences and productivity as men may receive lower salaries or positions (Goldin and Rouse, 2000; Black and Strahan,

[6] For a discussion of homophily, see McPherson et al. (2001).

2001; Sarsons et al., 2021).

We make two contributions to this literature. First, we examine gender differences in team size and composition in a setting where men and women have the same opportunities to select their preferred teams. Our findings show that women choose more connections than men when there are no costs to forming links, and both genders have the same opportunities to select team members. This result contrasts with findings in real-world environments, where women have fewer collaborators in fields such as economics, sociology, and computer science, as well as in collaborations within private companies like Enron (Lindenlaub and Prummer, 2020; Ductor et al., 2023). This discrepancy suggests that the lower number of collaborators observed among women in different workplaces may be due to a more hostile or challenging environment (Mengel, 2020; Sarsons et al., 2021; Wu, 2020; Hengel, 2022) or to gender differences in soft skills.[7]

Second, we study the role of gender stereotypes and gender homophily in team selection. Previous literature has documented gender homophily in several fields, among economists (Ductor and Prummer, 2024), among friends (Jackson et al., 2023; Brañas-Garza et al., 2022), in job search networks (Torres and Huffman, 2002; Zhu, 2022) and in the lab (Currarini and Mengel, 2016; Mengel, 2020). Observational studies face the challenge of distinguishing whether same-gender collaboration results from differences in interaction opportunities (e.g., women having fewer chances to establish collaborations) or from taste-based preferences. We contribute to their work by examining the sources of gender homophily in an experimental setting where men and women have equal interaction opportunities. To our knowledge, Currarini and Mengel (2016) is the only previous study that seeks to identify the sources of homophily within a laboratory setting. Their study focuses on homophily based on the declared willingness to pay to support an in-group, where groups are defined abstractly as Blue and Red. They do not study homophily in individual characteristics of

[7] Recent research suggests that women in certain professions, such as academia, face a more adverse environment compared to their male colleagues. For example, Mengel et al. (2019) finds that female economists receive, on average, lower teaching evaluations, which can impact their career progression. In addition, Sarsons et al. (2021) shows that female economists receive less credit for work done jointly with female coauthors, Wu (2020) highlights misogyny on the Econ Job Market Rumors website, and Hengel (2022) argues that women face discrimination in the publishing process, resulting in more time-intensive revisions. Our results contribute significantly to this literature by demonstrating that women have a preference for having more collaborators.

the participants (e.g. ethnicity, gender, or religion) to discard stereotypes as a source of homophily in attributes.

In contrast, our *team experiment* explicitly focuses on gender homophily and evaluates the role of stereotypes in explaining it. Specifically, participants choose their team's size and gender composition, and their payment bonuses depend on the performance of the team members. As participants are not required to interact with others to complete the task, any observed gender homophily is primarily attributable to taste-based preferences rather than beliefs about soft skills, gender differences in altruism, or motivations to mitigate team inefficiencies. Additionally, to evaluate the importance of gender stereotypes, we include an information treatment in which participants are shown performance distributions by gender and task.[8] Our findings show that gender homophily is an important driver of team formation, as observed in Currarini and Mengel (2016), and is primarily driven by taste-based preferences rather than performance considerations. Moreover, we find that providing information affects team composition: the proportion of women selected increases in the Emotion task (where women outperform men) and decreases in the Rotation task (where top performers are predominantly men). While individuals respond rationally to performance-related information, consistent with Gallen and Wasserman (2023), they do not fully adjust for their gender homophily preferences in our setting.

In experimental economics, several studies have analyzed the effect of gender stereotypes in competitive environments. A key finding in this literature is that the perception of whether men or women have an advantage on a task is crucial to explain gender differences in performance and on the willingness to compete (Iriberri and Rey-Biel, 2017; Halladay and Landsman, 2022; Geraldes et al., 2020; Hernandez-Arenaz, 2020; Bordalo et al., 2019). Our contribution to this literature is two-fold. On the one hand, we adapt text and math tasks to capture the participants' verbal and analytical skills in difficult exercises. Specifically, participants solved mathematical exercises from Spanish university entry exams and answered text-based questions from the C1 Cambridge English exam. On the other hand, we study how stereotypes and taste-based preferences affect team composition in a non-competitive setting. For this purpose, we systematically measure the distribution of performance and beliefs on these tasks; eliciting the distribution provides us with a more accurate understanding

[8] Gallen and Wasserman (2023) and Reuben et al. (2014) also examine the role of information to alter choices in the team composition of groups.

of performance and expectations, particularly the concentration at the extremes.

2 Gender differences in task performance and expectations

2.1 Task experiment

At the beginning of the task experiment, participants were asked to provide their Prolific ID, age, and gender. Based on the latter, each participant was assigned either a male or female avatar, following the procedures described in Mengel (2020). The participants were clearly informed that the avatar only represented their revealed gender identity and did not represent any other personal characteristics. For those who chose not to disclose their gender, a gender-neutral avatar, depicted as a black silhouette, was assigned.

After being assigned an avatar, participants were asked to complete a multiple choice questionnaire for 15 minutes. We employed a between-subjects design, where each participant was randomly assigned to one of four different gender-stereotyped tasks. Appendix C provides an example of each task. Here, we describe the tasks assigned to our participants:

- Text (N =186, Mean age = 25.7, 45.7% females). Participants were presented with five paragraphs, each containing three blank spaces. For each blank space, they were given three word options. Their task was to choose the word that best fit each blank space.

- Math (N =172, Mean age = 24.9, 53.5% females). Participants had to solve numerical problems and equations. For each question, participants were presented with three possible answers (only one was correct).

- Emotion (N =171, Mean age = 25.7, 52.1% females). Participants observed nine faces expressing different emotions (e.g., anger, happiness, etc). They were then asked to recognize which faces reflected a particular emotion.

- Rotation (N = 177, Mean age = 25.6, 49.2% females). Participants observed a 3D figure and three other figures. They had to identify which of these three figures was the same as or different from the original 3D figure shown.

Overall, 706 participants participated in this experiment. After completing the task, participants were asked to complete a brief questionnaire to collect socio-demographic information, including level of education, field of study, and political orientation. We also elicited their beliefs about the distribution of task performance among participants, distinguishing the performance of female and male participants, as well as a discrete measure of who they thought performed better in this task, males or females. [9]

At the end of the *task experiment*, we paid each correct answer at £0.30. In addition, participants received a completion fee of £1.50. The average duration of the experiment was 23.5 minutes (median: 18.1 minutes), and participants received on average £2.8 (median: £2.7).[10]

2.2 Belief experiment

Participants in the belief experiment (hereafter, observers) were randomly assigned two tasks from a set of four (Text, Math, Emotion or Rotation). For each task, observers were shown several examples, first without the answer and then with the correct answer. Observers were instructed not to complete the tasks themselves, but instead to imagine 100 male/female participants who had completed the tasks in a previous experiment.

To elicit the expectations of observers, we explained that participants in the task experiment could score as low as zero (if all answers were incorrect) and as high as 100 (if all answers were correct). We then asked: *In your opinion, how many men/women do you think got the score described below?*. The response options for this question were: 20% or less correct answers, 21-40%, 41-60%, 61-80%, and 81% or more correct answers.[11]

At the end of the belief experiment, one task and one gender were randomly selected for

[9] In total, 716 participants completed the task and received the payment for their participation, but 10 of them reported "other" when specifying their gender. Due to the small number of observations in this category, these participants were excluded from the subsequent analysis. Table A.1 of Appendix A describes our data, pooling the samples from our three experiments. By design, 50% of participants are female. The majority of male and female participants are between 18-45 years old and highly educated, with 62% holding an undergraduate or graduate degree.

[10] Only 15 participants spent less than eight minutes in the session. When we exclude these participants from the analysis, the results are quantitatively similar.

[11] We randomized the order of questions when asking for the performance of men and women in the task experiment.

payment. Observers earned £0.60 for each category they accurately predicted. However, £0.02 was deducted for every point of absolute discrepancy between the reported and actual percentages within each category, with a threshold of 25 points determining whether payment would be awarded for a particular category. This payment rule was explained to observers before starting the experiment.

Overall, 567 observers participated in this experiment. In the *belief experiment*, participants received a fixed payment of £1.60. In addition, they were paid a bonus (up to £3.00) depending on the accuracy of their responses. The median duration of the experiment was 14 minutes, and observers received on average £3.20. The experiment was conducted in July-August 2021 and April 2024.

2.3 Pre-registered hypotheses

Our female-typical tasks (Text and Emotion) are commonly associated with the stereotype that women excel over men, attributed to their perceived stronger verbal and emotional intelligence skills. These beliefs stem from cultural perceptions of women as being more expressive, empathetic, and skilled in verbal interactions compared to men, as well as having higher emotional intelligence, which is thought to enhance their ability to recognize and interpret emotional cues (Brody, 1997; Helgeson, 2020). Our male-typical tasks involve mathematical skills (Correll, 2001; Rudman and Kilianski, 2001; Kiefer and Sekaquaptewa, 2007; Bordalo et al., 2019) and mental rotation (Vandenberg and Kuse, 1978; Maeda and Yoon, 2013). In these tasks, there is the stereotype that men excel over women (Hyde et al., 2008; Hyde and Mertz, 2009; Iriberri and Rey-Biel, 2019; Dossi et al., 2021; Lambrecht et al., 2014; Kessels et al., 2014) because of the idea that men are more logical, competitive, and less affected by social or emotional influences when solving complex problems. However, stereotypes often do not match with actual performance; e.g., Boschini et al. (2012) do not find gender differences in performance in Text. Based on the previous literature on gender stereotypes and (expected) performance, we pre-registered the following hypotheses.[12]

[12] We pre-registered two additional hypotheses that will be addressed in future research. One examines the difference between the student sample and the Prolific sample. The other investigates the effect of incentives on belief elicitation, specifically i) whether rewarding observers affect their expectations on the performance of male and female participants in the belief experiment, and ii) whether the belief of observers (not subject to the task) differ from the beliefs of participants in the task experiment (who completed the task).

Hypothesis 1: *In the task experiment, female/male participants will answer more questions correctly in the Emotion/Math task. We expect no differences in performance of male and female participants in Text and Rotation.*

Hypothesis 2: *In the belief experiment, observers will expect male participants to perform better in the Math and Rotation tasks, while expecting female participants to excel in the Text and Emotion tasks.*

2.4 Results of task and belief experiment

2.4.1 Task experiment

Table 1 presents the performance of the participants in the *task experiment* by task and gender. It shows the percentage of men and women classified as low performers who score less than 20% correct answers in the task (i.e., 3 or less correct answers), medium performers who scored between 40-60% (i.e., 7-9 correct answers), and top performers with more than 80% (i.e, 13 or more correct answers).[13]

Table 1: Performance in each task by gender

Percentage	Text		Math		Emotion		Rotation	
of correct answers	Female	Male	Female	Male	Female	Male	Female	Male
Less or equal to 20%	5.9	11.9	5.4	7.5	0.0	0.0	3.5	2.2
Between 41-60%	40.0	29.7	45.7	28.8	24.7	26.8	23.0	14.4
More than 80%	4.7	7.9	6.5	13.8	11.2	7.3	41.4	53.3
Mean performance	7.55	7.39	8.42	8.15	10.74	10.23	10.89	11.20
T-test $p-$value	0.707		0.558		**0.040**		0.554	
Epps & Singleton $p-$value	0.527		**0.032**		0.206		**0.065**	

Note: The total number of participants for this experiment is 706 with equal split by gender.

We find no gender differences in average performance in the Text, Math, and Rotation tasks at any common significance level. For the Emotion task, there are gender differences

[13] This information is provided to participants as part of the information treatment in the *team experiment*, as detailed below.

in the expected direction at the 5% level (T, $p = 0.04$). These differences are driven by a higher concentration of top performer women (11.2% vs. 7.3%).

We also find that the distributions of performance in the math and rotation tasks are different between men and women (ES, $p = 0.032$ and $p = 0.065$ respectively). For the Math task, we observe a higher concentration of men in both tails of the distribution (low tail: 7.5% vs. 5.4%; high tail: 13.8% vs. 6.5%). In the Rotation task, men are also more concentrated in the top tail of the distribution (53.3% vs. 41.4%). We do not find differences for the other distributions of performance.

2.4.2 Belief experiment

We present the expected performance reported by observers in Table 2. Consistent with our pre-registered hypotheses, women are expected to outperform men in the Text (T, $p = 0.002$) and Emotion (T, $p < 0.001$) tasks. Using Epps-Singleton tests, we also observe gender differences in the distribution of expected performance for the Text (ES, $p = 0.027$) and Emotion tasks (ES, $p = 0.001$), with a higher concentration of women expected to be top performers. While the expected performance aligns with the observed performance in the Emotion task, as shown in Table 1, expectations diverge from observed performance in the remaining tasks.

Table 2: Expected task performance of females and males in each task

All participants

	Text		Math		Emotion		Rotation	
	Female	Male	Female	Male	Female	Male	Female	Male
Less or equal to 20%	7.9	10.1	12.7	12.6	5.0	8.6	10.0	10.3
Between 41-60%	18.6	20.4	20.3	19.7	16.1	17.5	20.2	19.2
More than 80%	35.1	29.9	28.3	28.4	44.8	36.5	34.4	34.4
Mean performance	9.65	9.10	8.78	8.80	10.39	9.64	9.32	9.30
T-test $p-$value	**0.002**		0.906		**0.000**		0.901	
Epps & Singleton $p-$value	**0.027**		0.987		**0.001**		0.959	

Note: The total number of participants for this experiment is 567 with equal split by gender.

When splitting the sample between female and male observers, see Table A.3, we find that both male and female observers expect differences in average performance in the Emotion task (T, $p < 0.001$ and $p = 0.077$, respectively). Specifically, observers generally expect higher average performance from females and a higher concentration of females among top performers. For the Text task, there are significant differences in the average expected performance reported by female and male observers (T, $p = 0.009$ and $p = 0.067$, respectively) where females are expected to outperform men.

In contrast, the expected performance reported by both female and male observers is very similar for the Math and Rotation tasks. Female observers indicate a slight concentration of females among top performers in Math and Rotation, whereas male observers report a slight concentration of males as top performers in these tasks. However, these differences are not statistically significant.

We therefore conclude that, consistent with the average performance in the Emotion task, both female and male observers correctly predict that women, on average, are better than men in this task. However, contrary to the observed performance, males and female observers expect women to outperform men, on average, in the Text task as well. For Text and Emotion, observers anticipate a higher concentration of women among top performers and of men among low performers. This expectation aligns with observed performance only for the Emotion task. Consistent with task performance, no significant differences in expected performance are found for the Math and Rotation tasks. To elicit expected performance, we asked observers to estimate the number of women and men, separately, that they expect to fall within five partitions of correct answers. See Figure C5 for an example of this elicitation. To calculate an average expected performance reported by observers, we use the corresponding midpoint of each category (i.e., 20% or less of correct answers, 21-40%, 41-60%, 61-80% and more than 80%. A more robust approach to test for gender differences involves using the full distribution of responses reported by observers. Accordingly, we calculated the p-values for all observers in the *belief experiment* using a Seemingly Unrelated Regression (SUR) to capture the entire distribution and account for the correlation of responses across the three partitions: 20% or less, between 41-60% and above 80%. We then computed the likelihood ratio (LR) for beliefs about females and males (restricted model), separately. The likelihood ratios from the SUR models for females and males were added up (unrestricted model). Using the standard formula for the likelihood ratio test $LR = -2(\mathcal{L}_r - \mathcal{L}_{ur})$, we obtained the LR value for each task.[14]. To construct the p-values, we used a 5% significance level and 9 degrees of freedom. Our conclusions are consistent with those derived from T-tests for mean differences and Epps & Singleton tests for distribution differences: significant gender differences are found in the distributions for Text and Emotion tasks using this test (both p-values < 0.001).[15]

To summarize, the expected average performance broadly coincided with actual performance, with the exception of Text, where women were expected to do better than they did. For the distribution, men were overrepresented in the tails for Text and Math, but were not expected to be. This is significant because in the team formation experiment, the top and average performance are relevant for the different payment schemes.

3 Gender differences in team formation

3.1 Experimental design and procedures

Participants in the team experiment (hereafter, team leaders) were assigned a male or female avatar at the beginning of the experiment based on their self-identified gender. As in the task experiment, team leaders were then introduced to one of the tasks (Text, Math, Emotion, Rotation), which they had 15 minutes to complete.

Before taking their decisions, team leaders were shown 10 avatars (5 males and 5 females), corresponding to participants who completed the task experiment (see Figure 1). Using this screen, team leaders were asked to create a team by selecting from 1 to 5 avatars; i.e., each team had a minimum of 2 members and a maximum of 6 members, including the team leader.

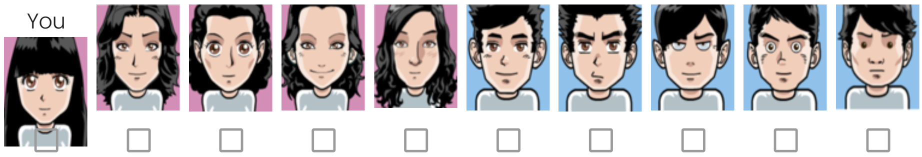
Team leaders formed three teams, each corresponding to one of the following payment rules (presented in random order):

- Minimum (Min): The payment for the team leader was determined by the performance of the team member with the lowest task score.

[14] The multiplication by two in this equation is needed to approximate the LR distribution to a chi-square, see Wooldridge (2013).

[15] We are grateful to Andreas Murr for suggesting this implementation.

Figure 1: Team formation experiment – Screen to select team members for female participants



- Maximum (Max): The payment for the team leader was determined by the performance of the team member with the highest task score.

- Median: The payment for the team leader was determined by the performance of the team member with the median task score.

When selecting their teams members for each payment rule, team leaders were informed that the avatars represented individuals who had previously completed the experiment individually. This approach was chosen to minimise the influence of beliefs about how men and women perform in groups on team composition; for example, women could have been selected for teams based on the expectation that they are more pro-social (Babcock et al., 2022; Brañas-Garza et al., 2018). Additionally, participants were allowed to freely select as many avatars as they wished (up to a maximum of 5), thereby allowing variation in both group sizes (degree) and group composition.[16]

Our between-subject experiment consisted of two treatments: Info and No-Info. In the Info treatment ($N = 796$), team leaders were shown the percentages of men and women who had scored less than or equal to 20% (bottom performers), between 41-60% (medium performers), and more than 80% correct answers (top performers) when working individually. These data corresponded to the actual performance of men and women in the *task experiment*. In the No-Info treatment ($N = 766$), placebo information was presented to team leaders; in particular, we showed them the average age of participants in the *task experiment*. As in the Info treatment, the information was disaggregated by gender and it included the average age of the youngest 20%, the average age of those in the middle 41-60%, and the

[16] Table A.4 shows the average number of times the participant selected a specific avatar. For females, we observe that the avatars appearing first are more likely to be selected (without considering themselves also referred as ”You”). For males, we observe the same pattern for male avatars, but a slight preference for the last female avatar.

average age of the oldest 20%.[17] By comparing the results between the Info and No-Info treatments, we can evaluate how information can mitigate the role of gender stereotypes in team selection.

Following the procedures of previous experiments, we elicited team leaders' expectations regarding task performance for females and males, separately. Although team members were exposed to the task, our conclusions about their beliefs remain consistent with those observed in the *belief experiment*, where observers were not exposed to the task. Overall, we find that people expect gender differences in both the average performance and distributions of performance for Text and Emotion tasks.[18] Thus, the beliefs elicited from our *belief experiment* provide a reliable basis for rationalizing the choices made by team leaders. In our team experiment, we also collected data on team leaders' level of education, areas of studies and political orientation.

To pay participants, we randomly selected one of the payment rules (Max, Median, Min). As in the task experiment each correct answer was paid at £0.30. This amount was added to their completion fee of £1.50. The average payment was £4.3. The median time the participants took to do the experiment was 25:53 minutes.

3.2 Pre-registered hypotheses

The first pre-registered hypothesis examines the effect of information about task performance on the team composition selected by team leaders. As noted in the previous section, this information corresponds to that displayed in Table 1 of Section 2.4.

Hypothesis 3: *Information vs Preferences on Team Composition. If the team composition chosen by participants remains the same, regardless of the treatment to which they were randomly assigned, it suggests that preferences are driving our results.*

The second hypothesis concerns the team composition selected for the four tasks (female-

[17] In the Info treatment there are $N = 205$ in Text, $N = 183$ in Math, $N = 203$ in Emotion and $N = 205$ in Rotation. In the No-Info treatment we have $N = 184$ participants in Text, $N = 194$ in Math, $N = 215$ in Emotion and $N = 173$ in Rotation. In total, we have 1,562 team leaders in this experiment with equal split by gender. In Table A.5, we present balance tests between Info and No-Info participants using socio-demographic characteristics. We do not find differences in the characteristics of our samples.

[18] Text: T, $p < 0.001$ and ES, $p < 0.001$; Math: T, $p = 0.993$ and ES, $p = 0.792$; Emotion: T, $p < 0.001$ and ES, $p < 0.001$; and Rotation: T, $p = 0.635$ and ES, $p = 0.649$.

type tasks: Text and Emotion, and male-type tasks: Math and Rotation). These hypotheses were informed by the findings from the *task* and *belief experiments*.

Hypothesis 4: *There will be a preference for selecting a higher proportion of female participants for female-type tasks (Text and Emotion), where females are expected to outperform males. We expect this preference to be particularly strong when team leaders receive information confirming that females perform better than males in the Emotion task (Info treatment). For the male-type tasks (Math and Rotation), we do not expect a preference for selecting a lower proportion of females in the No-Info treatment, as no differences in the expected performance between men and women were found in our belief experiment. However, in the Info treatment, we expect a lower proportion of female participants when team leaders receive information that males are less concentrated in the bottom performance and more concentrated in the top performance in the Rotation task.*

Our final hypothesis examines team size (out-degree) by contrasting the three types of payment rules.

Hypothesis 5: *On average, the number of team members will be higher when the team leader's payment is based on the performance of the team member with the highest task score and lower when it is based on the performance of the member with the lowest task score.*

3.3 Results of the team experiment

In this section, we test our hypotheses using Epps-Singleton (ES) tests to examine differences in distributions and Ordinary Least Squares (OLS) regressions without covariates, t-tests (T), to assess differences in means. First, we pool observations across all tasks to evaluate the effects of information on the team leaders' decisions about the proportion of females selected as team members under different payment rules (section 3.3.1). Next, we analyse the results by tasks and gender of team leaders (section 3.3.2). We finalize our result section analyzing information effects on the number of team members (section 3.3.3).

3.3.1 Information effects by payment rules

To test our *Hypothesis 3* on the effect of information, we pooled data from all four tasks and examined whether the proportion of females selected for the teams differed between the Info and No-Info treatments. Table 3 presents the results for the three payment rules.

Table 3: Proportion of females in the team by treatment

Pooling all tasks

Type of	Mean		T	ES	Observations	
Payment	No-Info	Info	p-value		No-Info	Info
Min	0.53	0.52	0.642	0.000	766	796
Med	0.52	0.54	0.199	0.000	766	796
Max	0.53	0.50	0.019	0.000	766	796

On average, the proportion of females under the maximum payment scheme in the Info treatment is 3pp smaller than the proportion reported by the No-Info group (T, $p = 0.019$). We do not find differences in means for the minimum (T, $p = 0.642$) and median payments (T, $p = 0.199$). Using ES tests we find significant differences in the distribution of the proportion of females selected by team leaders (ES, $p < 0.001$ for all minimum, maximum and median payments).

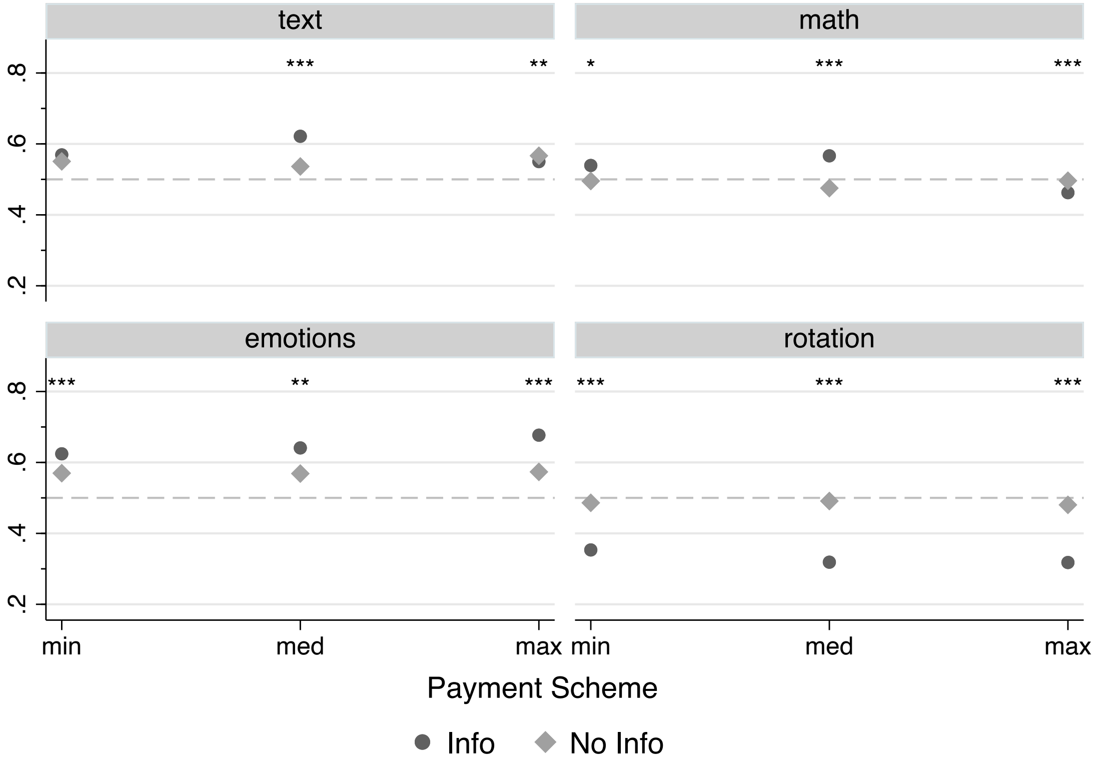
These results provide evidence in line with our *Hypothesis 3*, as we observe statistical differences in the distributions of the proportion of females for the three payment rules and differences in means under the maximum payment rule. However, the differences are quantitatively small, indicating that information has only a modest influence on participants' decisions; thus, preferences continue to play a dominant role in team formation.

3.3.2 Information effects by tasks and gender

To test *Hypothesis 4* for team composition, we analyse our results by task. Figure 2 displays the percentage of female participants selected by the team leaders in each task, distinguishing between the Info (●) and No-Info (⧫) treatments. The horizontal dashed line indicates gender parity (50-50) within the team.

Consistent with the findings of the *belief experiment*, which indicate that females are expected to outperform males in Text and Emotion tasks, we find that teams in the No-Info group predominantly consist of females (more than 50%) for these tasks. In contrast, for Math and Rotation tasks, we observe an approximately equal proportion of females and males in the teams (around 50%), aligning with the expected performances reported in the *belief experiment*.

Figure 2: Selection of the proportion of females by task and treatment



Note: ES p-values are reported as: *** $p < 0.01$, ** $p < 0.05$ and * $p < 0.10$

Do people react to information when forming their team? Participants who were assigned to the Info group respond to the information provided to them. For Emotion, the proportion of females is higher in the Info treatments, while the proportion of females is lower for Rotation. This pattern holds across all payment schemes and it is consistent with the information provided.[19] For the other two tasks (Text and Math), more females are selected under the Min and Median payment rules in the Info group, but this number decreases for both tasks under the maximum payment.

This result provides evidence supporting *Hypothesis 4*, as we observe that team leaders select a higher proportion of female members for female-type tasks (Text and Emotion) and an equal split for male-type tasks (Math and Rotation). Additionally, we confirm that information increases the proportion of females in Emotion when participants are exposed to the Info treatment and decreases it in Rotation.

[19] Recall that the information provided to the Info group corresponds to the main findings of the *task experiment*. These reveal a higher concentration of top-performing females in Emotion and top-performing males in Rotation.

Next, we examine whether team composition differs between male and female team leaders. Figures 3 and 4 illustrate the proportion of females, similar to Figure 2, with a dashed line indicating gender parity (50-50). We identify two main findings. The first relates to gender homophily, which can be observed in the No-Info treatment. When the team leader is a woman, the proportion of women in the team is consistently above 50% (ranging from 58% to 71%; see Table A.9 for the exact values). Conversely, in teams where the team leader is a man, the proportion of women in the No-Info treatment falls below 50% (ranging from 36% to 47%; see Table A.10 for the exact values). These results demonstrate that both male and female team leaders tend to prefer forming teams composed predominantly of members of their own gender, which we interpret as evidence of gender homophily.[20]

This is an important finding, as team leaders are not required to interact with others to complete the task; therefore, their preference cannot be attributed to beliefs about soft skills, gender differences in altruism, or incentives to shrink in teams.

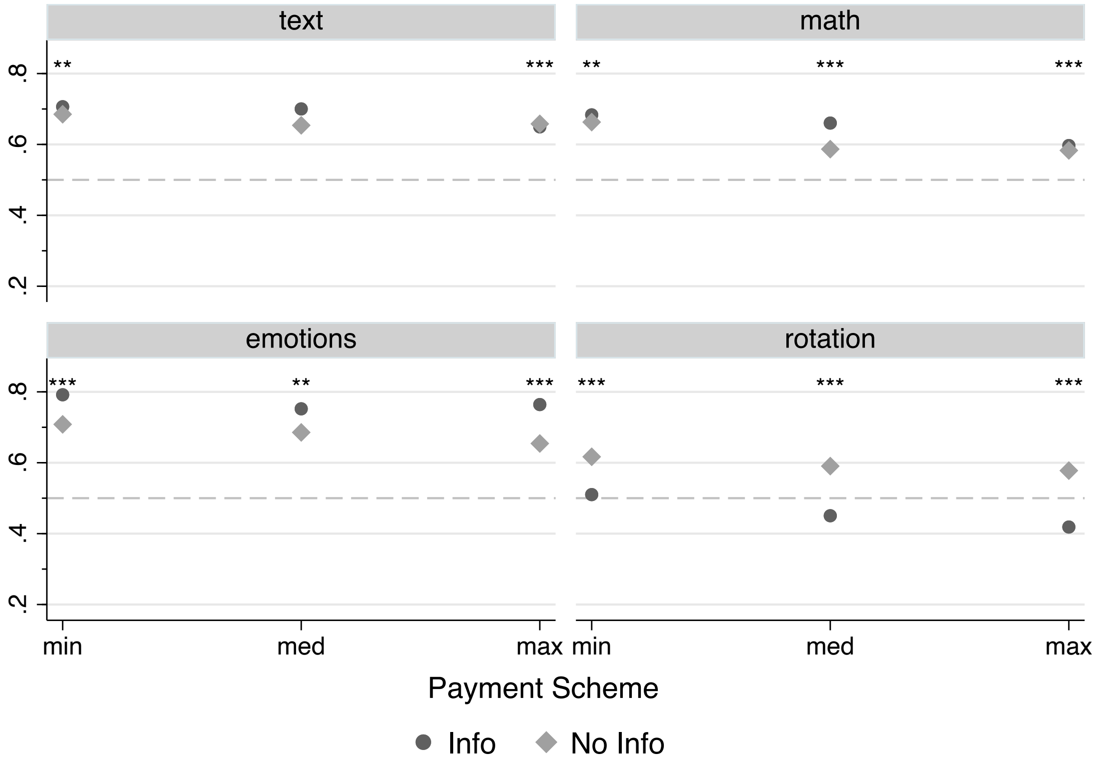
Second, we observe that information impacts men and women equally. Both men and women who receive information select a higher proportion of women in Emotion (for all three payment schemes) and a lower proportion of women for Rotation (for all three payment schemes), compared to the No-Info treatment. Thus, we find that information mitigates homophily, as individuals adjust their behavior based on the information provided. However, as mentioned previously, the effects of information are not quantitatively large. Clearly, homophily and, generally, preferences are a dominant force in team formation.

When exploring the heterogeneous effects of the Info treatment by gender, age, and education in Tables A.14-A.16, we find that females and males generally respond similarly to the information, with one exception: under the maximum payment rule (pooling all tasks), women exhibit a stronger reaction to the information treatment than men. No heterogeneous effects are observed on the basis of age or education.

[20] Our definition of homophily includes the team leader when computing the proportion of females in the team. This approach differs from the relative homophily index defined by Coleman (1958), which excludes the team leader from the calculations.

Figure 3: Proportion of females by task and treatment

Female team leaders



Note: ES p-values are reported as: *** $p < 0.01$, ** $p < 0.05$ and * $p < 0.10$

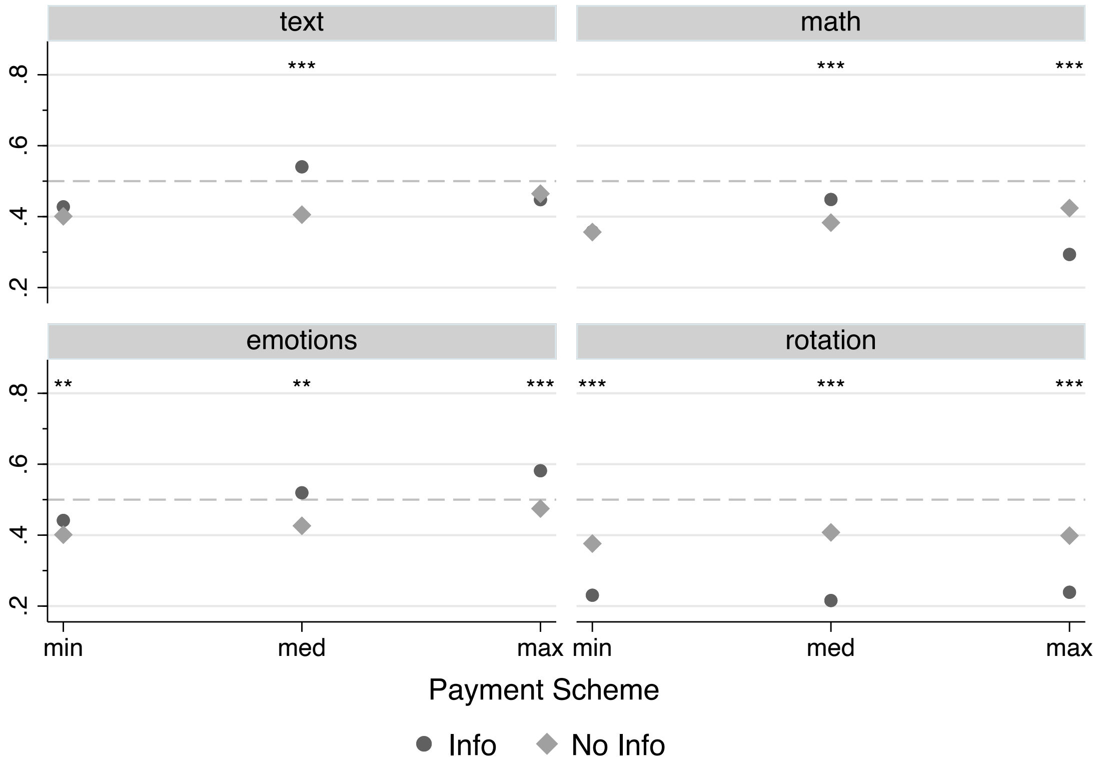
3.3.3 Information effects on team size

To examine the effect of the payment rules on team size, we pool again all tasks and examine the number of team members selected in the Info and the No-Info treatments. Overall, without differentiating between these two groups, we find that the number of members under the maximum payment is 1.2 team members larger than the number reported under the minimum payment (T, $p < 0.000$).

When examining differences in distributions between Info y No-Info across payments, only for the minimum payment, there are differences in distributions between Info and No-Info groups (EP, $p = 0.009$). For the median and maximum payment rules, we do not find significant differences in distributions (EP, $p = 0.164$ and $p = 0.203$). On average, the number of team members is larger for the participants allocated to the Info group under the minimum payment scheme by 0.23 members (T, $p = 0.002$) and under the median payment by 0.16 members (T, $p = 0.020$).

Figure 4: Proportion of females by task and treatment

Male team leaders



Note: ES p-values are reported as: *** $p < 0.01$, ** $p < 0.05$ and * $p < 0.10$

Table 4: Number of team members

Pooling all tasks

Type of	Mean		T	ES	Observations	
Payment	No-Info	Info	p-value		No-Info	Info
Min	3.41	3.64	0.002	0.009	766	796
Med	4.09	4.25	0.020	0.164	766	796
Max	4.78	4.68	0.158	0.203	766	796

Splitting the sample by gender of team leaders and pooling the Info and No-Info samples, we explore the number of team members selected by female and male team leaders. Table 5 shows that female team leaders tend to select slightly larger number of team members. This is an important finding that relates to differences in the degree of collaboration networks (Ductor et al., 2023). Our results highlight that in an environment where women and men have exactly the same costs and opportunities to collaborate with others, they select a higher number of team members than men. In Table 5, we also observe that the payment rule is important to determine the group size as more team members are selected in maximum payment rule than in the minimum. This result provides evidence supporting *Hypothesis 5*.

Table 5: Number of team members by the gender of the team leader

Type of	Task	Mean		ES	Observations	
Payment		Males	Females	p-value	Males	Females
Text	Min	3.49	3.53	0.201	188	201
	Med	4.19	4.14	0.545	188	201
	Max	4.80	4.61	0.038	188	201
Math	Min	3.31	3.73	0.060	187	190
	Med	4.12	4.42	0.007	187	190
	Max	4.79	4.78	0.246	187	190
Emotion	Min	3.30	3.58	0.014	194	224
	Med	4.09	4.14	0.363	194	224
	Max	4.69	4.71	0.008	194	224
Rotation	Min	3.48	3.82	0.012	209	169
	Med	4.01	4.29	0.002	209	169
	Max	4.60	4.86	0.128	209	169

4 Conclusion

This paper examines gender differences in team formation using data from three online experiments conducted across four distinct tasks (Text, Math, Emotion, and Rotation), which are associated with gender stereotypes. The findings provide evidence of gender homophily in team formation, whereby women (men) prefer to form teams predominantly composed of women (men). This preference persists regardless of the payment scheme used to incentivize team performance. Importantly, the observed homophily cannot be fully explained by gender stereotypes. Women expect to excel in female-typed tasks (Text and Emotion) but do not anticipate gender differences in performance for male-typed tasks (Math and Rotation). Similarly, men do not expect gender differences in performance across tasks, except for Rotation, where they still demonstrate a preference for teams with a majority of male members.

The paper also documents that gender homophily can be mitigated through the provision of information about actual performance. When team leaders are informed that women (men) outperform in Emotion (Rotation), the proportion of women (men) in teams increases, irrespective of the payment scheme, although the changes are quantitatively small. Additionally, the findings reveal that homophily is primarily driven by taste-based preferences rather than by gender stereotypes or statistical discrimination.

Our findings also reveal interesting insights into differences in team sizes, as female team leaders tend to form larger teams compared to their male counterparts. This contrasts with the smaller number of collaborators typically observed among women in co-authorship networks within fields such as economics, sociology, computer science, and even in private companies like Enron (Lindenlaub and Prummer, 2020), Ductor et al. (2023). We interpret this finding as evidence that women may face an adverse environment in these fields (Sarsons et al. (2021), Hengel (2022),Wu (2020)), which constrain their ability to form new collaborations. In an environment where men and women have equal opportunities to collaborate, as in our experiment, women would probably select a higher number of collaborators.

References

Adda, J., Dustmann, C., and Stevens, K. (2011). The Career Costs of Children.

Albanesi, S. and Olivetti, C. (2009). Production, Market Production and the Gender Wage Gap: Incentives and Expectations. *Review of Economic Dynamics*, 12(1):80–107.

Averett, S., Hoffman, S. D., et al. (2018). *The Oxford handbook of women and the economy.* Oxford University Press.

Azmat, G. and Petrongolo, B. (2014). Gender and the labor market: What have we learned from field and lab experiments? *Labour economics*, 30:32–40.

Babcock, L., Peyser, B., Vesterlund, L., and Weingart, L. (2022). *The No Club: putting a stop to women's dead-end work.* Simon and Schuster.

Bagues, M. F. and Esteve-Volart, B. (2010). Can gender parity break the glass ceiling? evidence from a repeated randomized experiment. *The Review of Economic Studies*, 77(4):1301–1328.

Bertrand, M. (2011). *New Perspectives on Gender*, volume 4 of *Handbook of Labor Economics*, chapter 17, pages 1543–1590. Elsevier.

Bertrand, M., Goldin, C., and Katz, L. F. (2010). Dynamics of the Gender Gap for Young Professionals in the Financial and Corporate Sectors. *American Economic Journal: Applied Economics*, pages 228–255.

Beugnot, J., Fortin, B., Lacroix, G., and Villeval, M. C. (2019). Gender and peer effects on performance in social networks. *European Economic Review*, 113:207–224.

Black, S. E. and Strahan, P. E. (2001). The division of spoils: rent-sharing and discrimination in a regulated industry. *American Economic Review*, pages 814–831.

Blau, F. D. and Kahn, L. M. (2000). Gender differences in pay. Technical report, National bureau of economic research.

Blau, F. D. and Kahn, L. M. (2016). The gender wage gap: Extent, trends, and explanations. Technical report, National Bureau of Economic Research.

Bordalo, P., Coffman, K., Gennaioli, N., and Shleifer, A. (2019). Beliefs about gender. *American Economic Review*, 109(3):739–773.

Boschini, A., Muren, A., and Persson, M. (2012). Constructing gender differences in the economics lab. *Journal of Economic Behavior & Organization*, 84(3):741–752.

Brañas-Garza, P., Capraro, V., and Rascon-Ramirez, E. (2018). Gender differences in altruism on mechanical turk: Expectations and actual behaviour. *Economics Letters*, 170:19–23.

Brañas-Garza, P., Ductor, L., and Kovářík, J. (2022). The role of unobservable characteristics in friendship network formation. *arXiv preprint arXiv:2206.13641*.

Brody, L. R. (1997). Gender and emotion: Beyond stereotypes. *Journal of Social issues*, 53(2):369–393.

Card, D., DellaVigna, S., Funk, P., and Iriberri, N. (2020). Are referees and editors in economics gender neutral? *The Quarterly Journal of Economics*, 135(1):269–327.

Charness, G. and Kuhn, P. (2010). Lab labor: What can labor economists learn from the lab? *Handbook of Labor Economics*, 4:229–330.

Charness, G. and Rustichini, A. (2011). Gender differences in cooperation with group membership. *Games and Economic Behavior*, 72(1):77–85.

Coleman, J. (1958). Relational analysis: the study of social organizations with survey methods. *Human organization*, 17(4):28–36.

Correll, S. J. (2001). Gender and the career choice process: The role of biased self-assessments. *American Journal of Sociology*, 106(6):1691–1730.

Croson, R. and Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, 47(2):448.

Currarini, S. and Mengel, F. (2016). Identity, homophily and in-group bias. *European Economic Review*, 90:40–55.

Dossi, G., Figlio, D., Giuliano, P., and Sapienza, P. (2021). Born in the family: Preferences for boys and the gender gap in math. *Journal of Economic Behavior & Organization*, 183:175–188.

Ductor, L., Goyal, S., and Prummer, A. (2018). Gender & collaboration. Technical report, Working Paper, School of Economics and Finance, Queen Mary University of London.

Ductor, L., Goyal, S., and Prummer, A. (2023). Gender and collaboration. *Review of Economics and Statistics*, 105(6):1366–1378.

Ductor, L. and Prummer, A. (2024). Gender homophily, collaboration, and output. *Journal of Economic Behavior & Organization*, 221:477–492.

Eckel, C. and Grossman, P. (2008). Men, Women and Risk Aversion: Experimental Evidence. *Handbook of experimental economics results*, 1:1061–1073.

Fenoll, A. A. and Zaccagni, S. (2022). Gender mix and team performance: Differences between exogenously and endogenously formed teams. *Labour Economics*, 79:102269.

Gallen, Y. and Wasserman, M. (2023). Does information affect homophily? *Journal of Public Economics*, 222:104876.

Geraldes, D., Riedl, A., and Strobel, M. (2020). Gender differences in performance under competition: Is there a stereotype threat shadow? *Available at SSRN 3754872*.

Goldin, C., Katz, L. F., and Kuziemko, I. (2006). The homecoming of american college women: The reversal of the college gender gap. *Journal of Economic Perspectives*, 20(4):133–156.

Goldin, C. and Rouse, C. (2000). Orchestrating impartiality: The impact of "blind" auditions on female musicians. *The American Economic Review*, 90(4):715–741.

Halladay, B. and Landsman, R. (2022). Perception matters: The role of task gender stereotype on confidence and tournament selection. *Journal of Economic Behavior & Organization*, 199:35–43.

Helgeson, V. S. (2020). *Psychology of gender*. Routledge.

Hengel, E. (2022). Publishing while female: Are women held to higher standards? evidence from peer review. *The Economic Journal*, 132(648):2951–2991.

Herbst, D. and Mas, A. (2015). Peer effects on worker output in the laboratory generalize to the field. *Science*, 350(6260):545–549.

Hernandez-Arenaz, I. (2020). Stereotypes and tournament self-selection: A theoretical and experimental approach. *European Economic Review*, 126:103448.

Hyde, J. S., Fennema, E., and Lamon, S. J. (1990). Meta-analysis and the psychology of gender differences. *Psychological Bulletin*, 107(2):297–326.

Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B., and Williams, C. C. (2008). Gender similarities characterize math performance. *Science*, 321(5888):494–495.

Hyde, J. S. and Mertz, J. E. (2009). Gender, culture, and mathematics performance. *Proceedings of the national academy of sciences*, 106(22):8801–8807.

Iriberri, N. and Rey-Biel, P. (2017). Stereotypes are only a threat when beliefs are reinforced: On the sensitivity of gender differences in performance under competition to information provision. *Journal of Economic Behavior & Organization*, 135:99–111.

Iriberri, N. and Rey-Biel, P. (2019). Competitive pressure widens the gender gap in performance: Evidence from a two-stage competition in mathematics. *The Economic Journal*, 129(620):1863–1893.

Jackson, M. O., Nei, S. M., Snowberg, E., and Yariv, L. (2023). The dynamics of networks and homophily. Technical report, National Bureau of Economic Research.

Kessels, R. P., Montagne, B., Hendriks, A. W., Perrett, D. I., and De Haan, E. H. (2014). Assessment of perception of morphed facial expressions using the emotion recognition task: Normative data from healthy participants aged 8–75. *Journal of neuropsychology*, 8(1):75–93.

Kiefer, A. K. and Sekaquaptewa, D. (2007). Implicit stereotypes, gender identification, and math-related outcomes: A prospective study of female college students. *Psychological Science*, 18(1):13–18.

Lalanne, M. and Seabright, P. (2016). The old boy network: The impact of professional networks on remuneration in top executive jobs.

Lambrecht, L., Kreifelts, B., and Wildgruber, D. (2014). Gender differences in emotion recognition: Impact of sensory modality and emotional category. *Cognition & emotion*, 28(3):452–469.

Lindenlaub, I. and Prummer, A. (2020). Network structure and performance. *The Economic Journal*.

Maeda, Y. and Yoon, S. Y. (2013). A meta-analysis on gender differences in mental rotation ability measured by the purdue spatial visualization tests: Visualization of rotations (psvt: R). *Educational Psychology Review*, 25:69–94.

McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444.

Mengel, F. (2020). Gender differences in networking. *The Economic Journal*, 130(630):1842–1873.

Mengel, F., Sauermann, J., and Zölitz, U. (2019). Gender bias in teaching evaluations. *Journal of the European economic association*, 17(2):535–566.

Niederle, M., Segal, C., and Vesterlund, L. (2013). Gender and competition. In *Handbook of Experimental Economics*, pages 127–184. Elsevier.

Niederle, M. and Vesterlund, L. (2007). Do women shy away from competition? do men compete too much? *Quarterly Journal of Economics*, 122(3):1067–1101.

Palan, S. and Schitter, C. (2018). Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27.

Reuben, E., Sapienza, P., and Zingales, L. (2014). How stereotypes impair women's careers in science. *Proceedings of the National Academy of Sciences*, 111(12):4403–4408.

Rudman, L. A. and Kilianski, S. E. (2001). Implicit and explicit attitudes toward female authority. *Personality and Social Psychology Bulletin*, 27(11):1325–1338.

Sarsons, H., Gërxhani, K., Reuben, E., and Schram, A. (2021). Gender differences in recognition for group work. *Journal of Political economy*, 129(1):101–147.

Torres, L. and Huffman, M. L. (2002). Social networks and job search outcomes among male and female professional, technical, and managerial workers. *Sociological Focus*, 35(1):25–42.

Vandenberg, S. G. and Kuse, A. R. (1978). Mental rotations, a group test of three-dimensional spatial visualization. *Perceptual and Motor Skills*, 47(2):599–604.

Wooldridge, J. M. (2013). *Introductory Econometrics: A Modern Approach 5th ed.*, chapter 17.

Wu, A. H. (2020). Gender bias among professionals: an identity-based interpretation. *Review of Economics and Statistics*, 102(5):867–880.

Zhu, M. (2022). Beyond the "old boys' network": Social networks and job finding at community colleges. *Journal of Human Resources.*

A Appendix: Tables

Table A.1: Summary Statistics: Task, belief and team experiments

	Mean	SD	min	max	obs
Female	0.50	0.50	0	1	2835
Under 25	0.41	0.49	0	1	2844
Btw 25-45	0.51	0.50	0	1	2844
Older than 45	0.08	0.27	0	1	2844
High School or lower	0.16	0.36	0	1	2815
Technical	0.18	0.39	0	1	2815
Undergraduate	0.34	0.47	0	1	2815
Postgraduate	0.28	0.45	0	1	2815

Table A.2: Comparison of elicited expectations about females and males
P-values of Epps-Singleton Tests

Percentage of correct answers	Text	Math	Emotion	Rotation
Less than 20%	0.031	0.661	0.001	0.800
Between 40-60%	0.524	0.806	0.306	0.886
More than 80%	0.121	0.998	0.004	0.984

Note: Table is constructed using the elicitation of expectations with incentives.

Table A.3: Expected performance of females and males in each task

By observer's gender

	Text		Math		Emotion		Rotation	
	Female	Male	Female	Male	Female	Male	Female	Male
				All observers				
Less or equal to 20%	7.9	10.1	12.7	12.6	5.0	8.6	10.0	10.3
Between 41-60%	18.6	20.4	20.3	19.7	16.1	17.5	20.2	19.2
More than 80%	35.1	29.9	28.3	28.4	44.8	36.5	34.4	34.4
Mean performance	9.65	9.10	8.78	8.80	10.39	9.64	9.32	9.30
T-test $p-$value	**0.002**		0.906		**0.000**		0.901	
Epps & Singleton $p-$value	**0.027**		0.987		**0.001**		0.959	
SUR: LR test $p-$value	**0.000**		0.999		**0.000**		0.827	
				Female observers				
Less or equal to 20%	7.7	11.0	13.3	12.5	4.8	9.6	10.3	11.8
Between 41-60%	17.8	20.4	20.6	19.3	15.4	17.8	20.0	18.6
More than 80%	38.1	32.1	27.6	27.1	46.6	36.9	34.6	33.5
Mean performance	9.82	9.12	8.68	8.69	10.51	9.49	9.28	9.17
T-test $p-$value	**0.009**		0.952		**0.000**		0.697	
Epps & Singleton $p-$value	0.114		0.885		**0.000**		0.651	
SUR: LR test $p-$value	**0.001**		0.602		**0.000**		0.053	
				Male observers				
Less or equal to 20%	8.1	9.3	12.0	12.6	5.3	7.6	9.7	9.0
Between 41-60%	19.5	20.5	20.0	20.2	16.9	17.2	20.4	19.8
More than 80%	32.1	27.6	29.0	29.7	43.1	36.1	34.2	35.2
Mean performance	9.48	9.08	8.88	8.91	10.28	9.78	9.36	9.41
T-test $p-$value	**0.067**		0.916		**0.077**		0.828	
Epps & Singleton $p-$value	0.444		0.923		0.315		0.870	
SUR: LR test $p-$value	0.639		0.820		**0.039**		0.289	

Note: The total number of participants for this experiment is 567 with equal split by gender.

Table A.4: Team Experiment: Selection of Avatars

Avatar	Obs	Mean	Std. Dev.	Min	Max
Selected by Female Team Leaders					
Female Avatar 1 (You)	784	3.00	0.00	3	3
Female Avatar 2	784	1.32	1.14	0	3
Female Avatar 3	784	1.32	1.03	0	3
Female Avatar 4	784	1.06	0.97	0	3
Female Avatar 5	784	1.13	1.04	0	3
Male Avatar 1	784	1.34	1.04	0	3
Male Avatar 2	784	1.09	1.01	0	3
Male Avatar 3	784	1.00	0.98	0	3
Male Avatar 4	784	0.71	0.90	0	3
Male Avatar 5	784	0.66	0.85	0	3
Selected by Male Team Leaders					
Female Avatar 1	778	0.92	1.07	0	3
Female Avatar 2	778	0.75	0.93	0	3
Female Avatar 3	778	1.06	0.98	0	3
Female Avatar 4	778	1.12	1.02	0	3
Female Avatar 5	778	1.35	1.10	0	3
Male Avatar 1 (You)	778	3.00	0.00	3	3
Male Avatar 2	778	1.36	1.09	0	3
Male Avatar 3	778	1.22	1.00	0	3
Male Avatar 4	778	0.75	0.89	0	3
Male Avatar 5	778	0.68	0.88	0	3

Note: This table summarizes the statistics for the number of times each avatar was selected across all three payment schemes. Figure 1 illustrates the sequence associated with each avatar number.

Table A.5: Balance Tests: Info vs No-Info

				Observations	
	No-Info	Info	p-value	No-Info	Info
Female	0.50	0.51	0.8026	766	796
Under 25	0.32	0.32	0.7288	766	796
Btw 25-45	0.59	0.57	0.3202	766	796
Older than 45	0.09	0.11	0.2727	766	796
High School or lower	0.23	0.23	0.9629	765	795
Technical	0.09	0.10	0.4432	765	795
Undergraduate	0.39	0.39	0.9269	765	795
Posgraduate	0.29	0.28	0.7151	765	795

Table A.6: Number of Team Members – Pooling Tasks and By Task

Team Members	**Minimum**			**Median**			**Maximum**		
	All tasks								
	Male	Female	Total	Male	Female	Total	Male	Female	Total
2	41.0	29.6	35.3	16.1	9.8	12.9	11.7	7.3	9.5
3	19.3	21.4	20.4	24.9	24.1	24.5	14.1	15.4	14.8
4	14.1	19.3	16.7	18.1	23.2	20.7	14.0	17.4	15.7
5	10.2	13.4	11.8	14.7	17.9	16.3	11.2	16.5	13.8
6	15.4	16.3	15.9	26.2	25.0	25.6	49.0	43.5	46.2
Chi-2 $p-$value		0.000			0.001			0.000	
	Text								
	Male	Female	Total	Male	Female	Total	Male	Female	Total
2	38.3	33.8	36.0	13.3	11.4	12.3	9.6	8.5	9.0
3	20.2	18.9	19.5	25.5	26.4	26.0	13.8	16.9	15.4
4	13.8	21.4	17.7	18.6	20.9	19.8	15.4	18.9	17.2
5	9.0	11.9	10.5	14.4	18.9	16.7	9.0	16.9	13.1
6	18.6	13.9	16.2	28.2	22.4	25.2	52.1	38.8	45.2
Chi-2 $p-$value		0.211			0.549			0.044	
	Math								
	Male	Female	Total	Male	Female	Total	Male	Female	Total
2	41.2	27.9	34.5	16.6	5.8	11.1	11.8	7.4	9.6
3	20.3	22.1	21.2	24.1	22.1	23.1	12.3	14.2	13.3
4	17.7	19.0	18.3	17.1	24.7	21.0	12.8	16.8	14.9
5	8.0	11.6	9.8	15.0	19.0	17.0	11.2	15.8	13.5
6	12.8	19.5	16.2	27.3	28.4	27.9	51.9	45.8	48.8
Chi-2 $p-$value		0.065			0.009			0.253	
	Emotion								
	Male	Female	Total	Male	Female	Total	Male	Female	Total
2	44.9	31.3	37.6	13.4	12.1	12.7	11.9	6.7	9.1
3	19.6	22.3	21.1	27.8	26.3	27.0	17.5	18.3	17.9
4	10.8	16.1	13.6	20.6	19.6	20.1	11.3	13.8	12.7
5	10.3	17.9	14.4	12.4	19.6	16.3	8.8	19.2	14.4
6	14.4	12.5	13.4	25.8	22.3	23.9	50.5	42.0	45.9
Chi-2 $p-$value		0.017			0.382			0.011	
	Rotation								
	Male	Female	Total	Male	Female	Total	Male	Female	Total
2	39.7	24.3	32.8	20.6	9.5	15.6	13.4	6.5	10.3
3	17.2	22.5	19.6	22.5	20.7	21.7	12.9	11.2	12.2
4	14.4	21.3	17.5	16.3	29.0	22.0	16.3	20.7	18.3
5	12.9	11.2	12.2	16.8	13.0	15.1	15.3	13.0	14.3
6	15.8	20.7	18.0	23.9	27.8	25.7	42.1	48.5	45.0
Chi-2 $p-$value		0.016			0.003			0.151	

Table A.7: Mean number of team members by task and treatment.

		Mean			Observations	
Task	Payment	No-Info	Info	p-value	No-Info	Info
Text	Max	4.69	4.71	1.000	184	205
	Min	3.36	3.65	0.039	184	205
	Med	4.00	4.31	0.025	184	205
Math	Max	4.95	4.61	0.007	194	183
	Min	3.51	3.54	0.782	194	183
	Med	4.17	4.38	0.151	194	183
Emotion	Max	4.78	4.62	0.240	215	203
	Min	3.27	3.64	0.012	215	203
	Med	4.04	4.20	0.258	215	203
Rotation	Max	4.67	4.75	0.586	173	205
	Min	3.52	3.72	0.190	173	205
	Med	4.14	4.13	0.880	173	205

Table A.8: Proportion of females in the team by task and treatment

Type of	Task	Mean		ES	Observations	
Payment		No-Info	Info	p-value	No-Info	Info
Text	Min	0.55	0.57	0.294	184	205
	Med	0.54	0.62	0.000	184	205
	Max	0.57	0.55	0.012	184	205
Math	Min	0.50	0.54	0.078	194	183
	Med	0.48	0.57	0.000	194	183
	Max	0.50	0.46	0.000	194	183
Emotion	Min	0.57	0.62	0.001	215	203
	Med	0.57	0.64	0.012	215	203
	Max	0.57	0.68	0.000	215	203
Rotation	Min	0.49	0.35	0.000	173	205
	Med	0.49	0.32	0.000	173	205
	Max	0.48	0.32	0.000	173	205

Table A.9: Female selection of the proportion of females in the team by task and treatment

Type of	Task	Mean		ES	Observations	
Payment		No-Info	Info	p-value	No-Info	Info
Text	Min	0.69	0.71	0.045	97	104
	Med	0.65	0.70	0.107	97	104
	Max	0.66	0.65	0.009	97	104
Math	Min	0.66	0.68	0.012	88	102
	Med	0.59	0.66	0.003	88	102
	Max	0.58	0.60	0.002	88	102
Emotion	Min	0.71	0.79	0.006	118	106
	Med	0.69	0.75	0.044	118	106
	Max	0.65	0.76	0.000	118	106
Rotation	Min	0.62	0.51	0.000	79	90
	Med	0.59	0.45	0.000	79	90
	Max	0.58	0.42	0.000	79	90

Table A.10: Male selection of the proportion of females in the team by task and treatment

Type of	Task	Mean		ES	Observations	
Payment		No-Info	Info	p-value	No-Info	Info
Text	Min	0.40	0.43	0.574	87	101
	Med	0.41	0.54	0.000	87	101
	Max	0.46	0.45	0.112	87	101
Math	Min	0.36	0.36	0.272	106	81
	Med	0.38	0.45	0.001	106	81
	Max	0.42	0.29	0.000	106	81
Emotion	Min	0.40	0.44	0.031	97	97
	Med	0.43	0.52	0.016	97	97
	Max	0.47	0.58	0.000	97	97
Rotation	Min	0.38	0.23	0.000	94	115
	Med	0.41	0.22	0.000	94	115
	Max	0.40	0.24	0.000	94	115

Table A.11: Proportion of females in the team by task and treatment

Excluding the team leader from both the selection of team members and the overall count of team participants

Type of	Task	Mean		ES	Observations	
Payment		No-Info	Info	p-value	No-Info	Info
Text	Min	0.56	0.62	0.077	184	205
	Med	0.54	0.67	0.000	184	205
	Max	0.58	0.57	0.004	184	205
Math	Min	0.53	0.55	0.022	194	183
	Med	0.49	0.58	0.007	194	183
	Max	0.52	0.44	0.000	194	183
Emotion	Min	0.62	0.69	0.035	215	203
	Med	0.58	0.70	0.000	215	203
	Max	0.58	0.73	0.000	215	203
Rotation	Min	0.52	0.31	0.000	173	205
	Med	0.52	0.27	0.000	173	205
	Max	0.50	0.28	0.000	173	205

Table A.12: Female selection of the proportion of females in the team by task and treatment

Excluding the team leader from both the selection of female members and the overall count of team participants

Type of	Task	Mean		ES	Observations	
Payment		No-Info	Info	p-value	No-Info	Info
Text	Min	0.51	0.58	0.067	97	104
	Med	0.52	0.60	0.040	97	104
	Max	0.55	0.54	0.023	97	104
Math	Min	0.51	0.55	0.033	88	102
	Med	0.44	0.55	0.003	88	102
	Max	0.47	0.48	0.001	88	102
Emotion	Min	0.57	0.70	0.013	118	106
	Med	0.57	0.67	0.037	118	106
	Max	0.55	0.69	0.000	118	106
Rotation	Min	0.44	0.30	0.000	79	90
	Med	0.46	0.25	0.000	79	90
	Max	0.45	0.24	0.000	79	90

Table A.13: Male selection of the proportion of females in the team by task and treatment

Excluding the team leader from both the selection of male members and the overall count of team participants

Type of	Task	Mean		ES	Observations	
Payment		No-Info	Info	p-value	No-Info	Info
Text	Min	0.62	0.66	0.446	87	101
	Med	0.56	0.74	0.000	87	101
	Max	0.62	0.59	0.208	87	101
Math	Min	0.54	0.56	0.290	106	81
	Med	0.54	0.62	0.114	106	81
	Max	0.56	0.40	0.001	106	81
Emotion	Min	0.68	0.67	0.361	97	97
	Med	0.58	0.73	0.001	97	97
	Max	0.62	0.78	0.000	97	97
Rotation	Min	0.58	0.32	0.000	94	115
	Med	0.58	0.29	0.000	94	115
	Max	0.55	0.31	0.000	94	115

Table A.14: Heterogeneous effects by gender

	(1)	(2)	(3)
VARIABLES	min	med	max
Female (=1)	0.290***	0.230***	0.183***
	[0.016]	[0.014]	[0.013]
Info Treatment (=1)	-0.024	0.016	-0.052***
	[0.017]	[0.016]	[0.016]
Female*Info	0.030	-0.004	0.044**
	[0.024]	[0.022]	[0.022]
Observations	1,562	1,562	1,562
R-squared	0.287	0.215	0.186

Robust standard errors in brackets

*** p<0.01, ** p<0.05, * p<0.1

Table A.15: Heterogeneous effects by age

	(1)	(2)	(3)
VARIABLES	min	med	max
Info Treatment (=1)	-0.009	0.016	-0.040*
	[0.025]	[0.021]	[0.021]
Age: 25-45 (=1)	-0.016	0.020	-0.000
	[0.021]	[0.018]	[0.016]
Age: >45 (=1)	-0.065*	-0.058*	-0.041
	[0.035]	[0.034]	[0.031]
Age: 25-45*Info	-0.001	-0.006	0.018
	[0.032]	[0.027]	[0.027]
Age: >45*Info	0.031	0.043	0.018
	[0.051]	[0.047]	[0.045]
Observations	1,562	1,562	1,562
R-squared	0.003	0.006	0.006

Note: Reference category is under 25. Robust standard errors in brackets.

*** p<0.01, ** p<0.05, * p<0.1

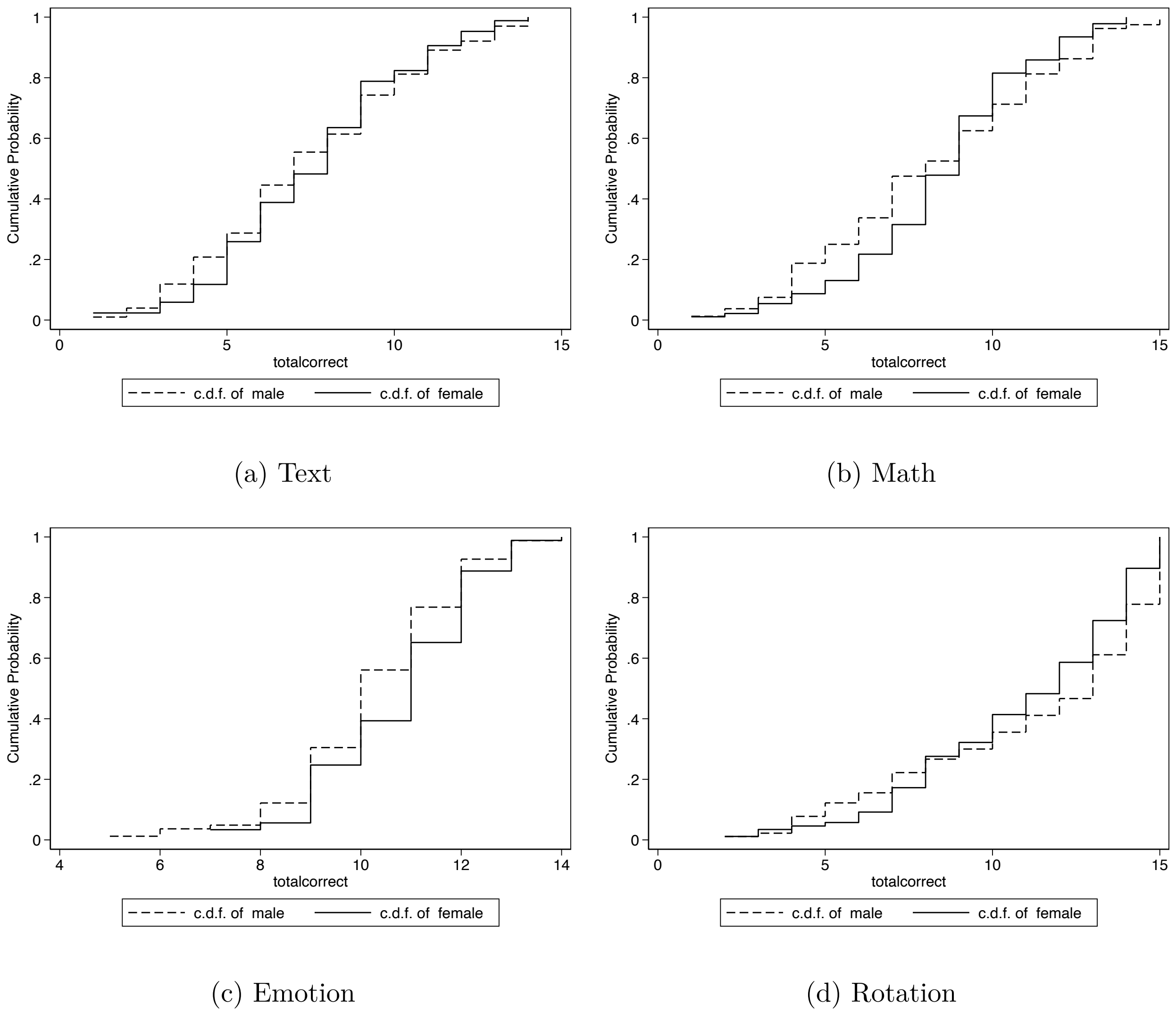
Table A.16: Heterogeneous effects by educational level

	(1)	(2)	(3)
Variables	min	med	max
Info Treatment (=1)	-0.032	0.006	-0.024
	[0.030]	[0.026]	[0.025]
Technical	-0.043	0.004	0.001
	[0.040]	[0.033]	[0.031]
Undergraduate	0.017	0.031	0.020
	[0.024]	[0.021]	[0.019]
Graduate	-0.013	0.021	0.009
	[0.027]	[0.023]	[0.020]
Technical*Info	0.008	-0.015	-0.033
	[0.058]	[0.049]	[0.048]
Undergraduate*Info	0.028	0.016	0.005
	[0.037]	[0.033]	[0.032]
Graduate*Info	0.051	0.021	-0.007
	[0.041]	[0.035]	[0.034]
Observations	1,560	1,560	1,560
R-squared	0.007	0.007	0.006

Note: Reference category is higher school or below. Robust standard errors in brackets.
*** p<0.01, ** p<0.05, * p<0.1

B Appendix: Figures

Figure B.1: Cumulative distribution of task performance (number of correct answers) by gender



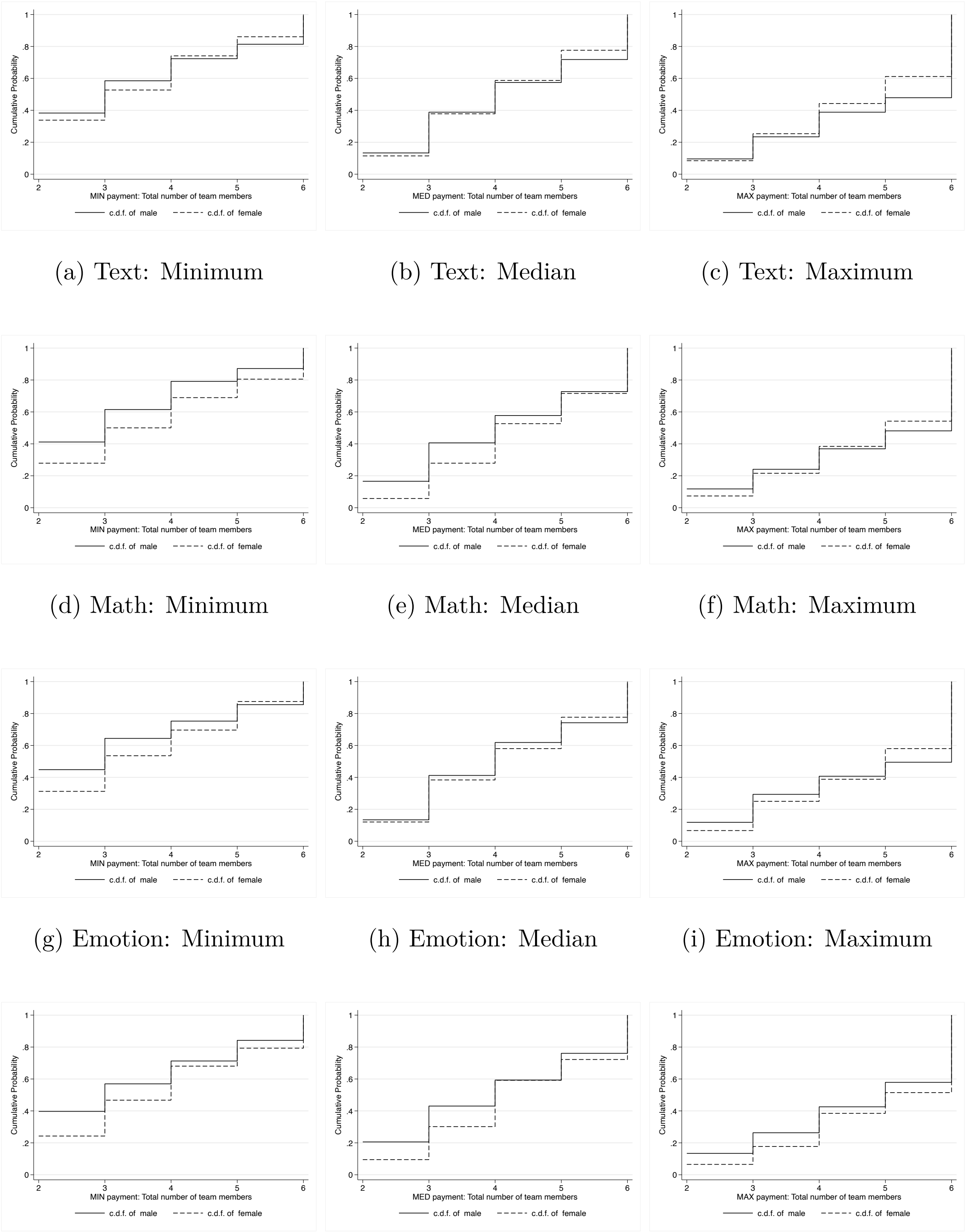
(a) Text

(b) Math

(c) Emotion

(d) Rotation

Figure B.2: Cumulative distributions of the number of team members by gender of team leader, task and type of payment



(a) Text: Minimum

(b) Text: Median

(c) Text: Maximum

(d) Math: Minimum

(e) Math: Median

(f) Math: Maximum

(g) Emotion: Minimum

(h) Emotion: Median

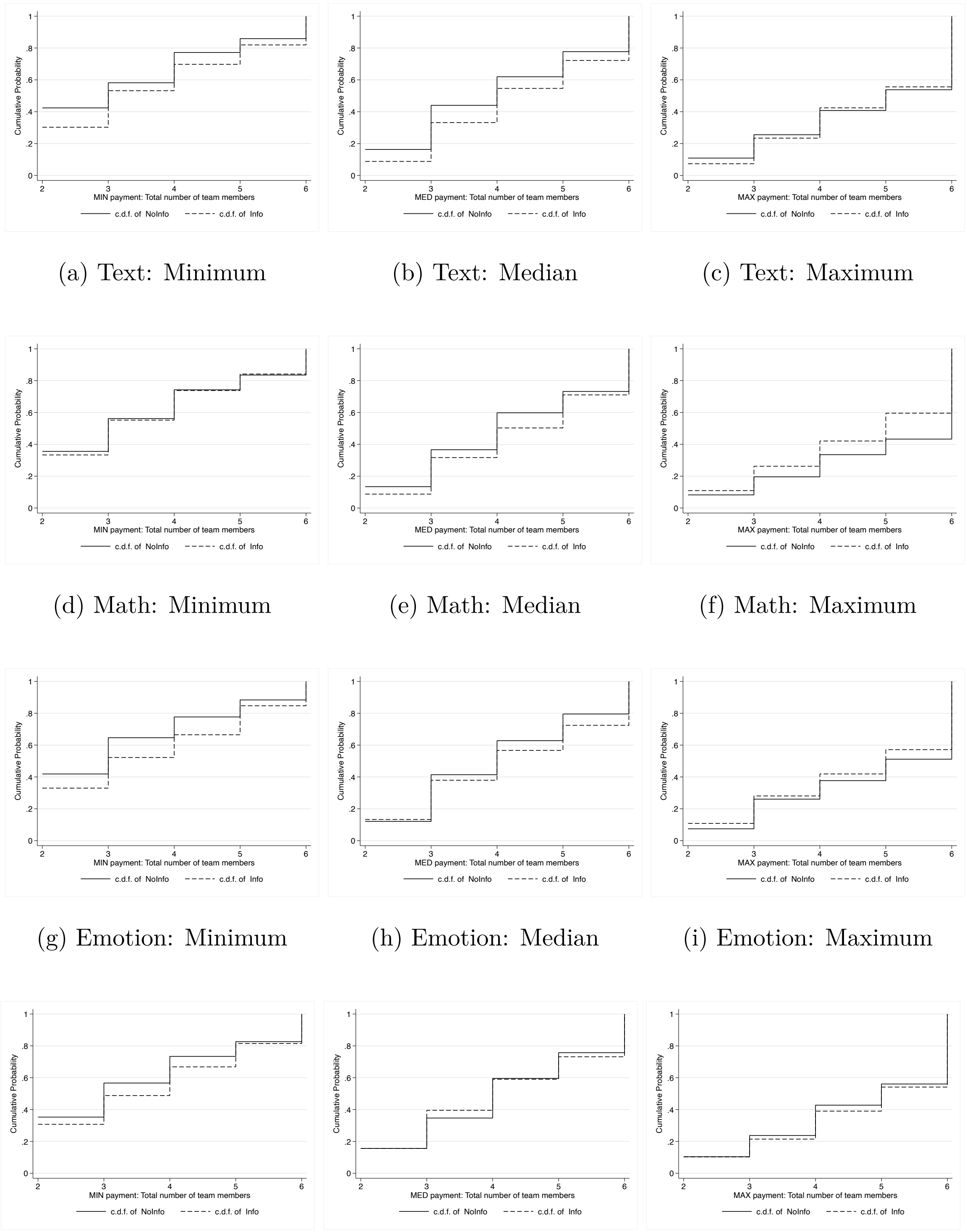
(i) Emotion: Maximum

(j) Rotation: Minimum

(k) Rotation: Median

(l) Rotation: Maximum

Figure B.3: Cumulative distributions of the number of team members by info treatment and type of payment



(a) Text: Minimum

(b) Text: Median

(c) Text: Maximum

(d) Math: Minimum

(e) Math: Median

(f) Math: Maximum

(g) Emotion: Minimum

(h) Emotion: Median

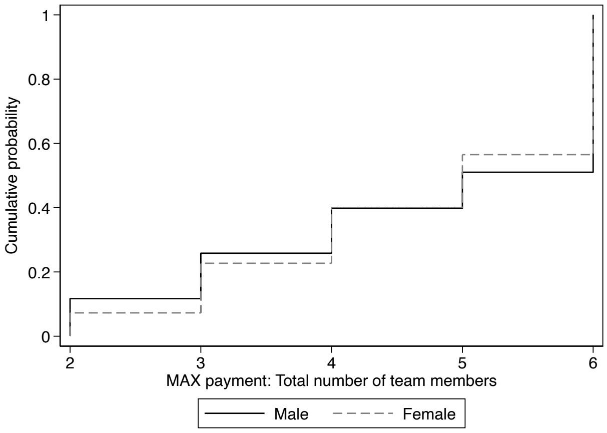
(i) Emotion: Maximum

(j) Rotation: Minimum

(k) Rotation: Median

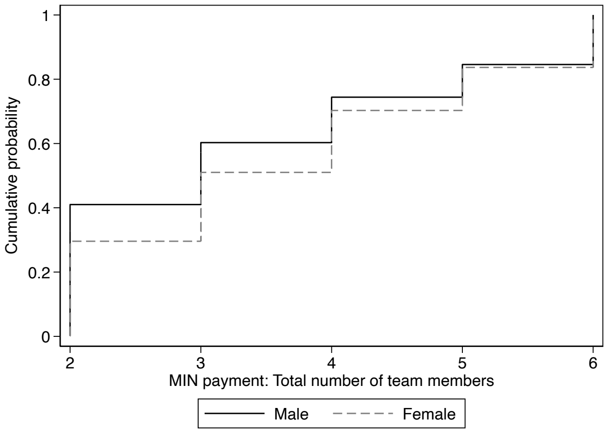
(l) Rotation: Maximum

Figure B.4: Payment based on the best performance: Number of team members by gender



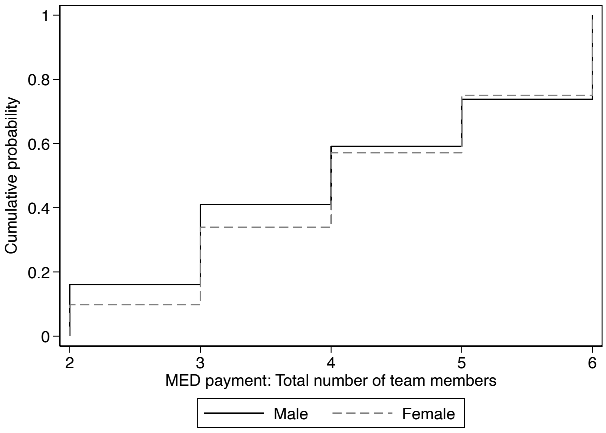
Note: P-value of two-sample Wilcoxon rank-sum (Mann-Whitney) test is 0.6319.

Figure B.5: Payment based on the worst performance: Number of team members by gender



Note: P-value of two-sample Wilcoxon rank-sum (Mann-Whitney) test is 0.0001.

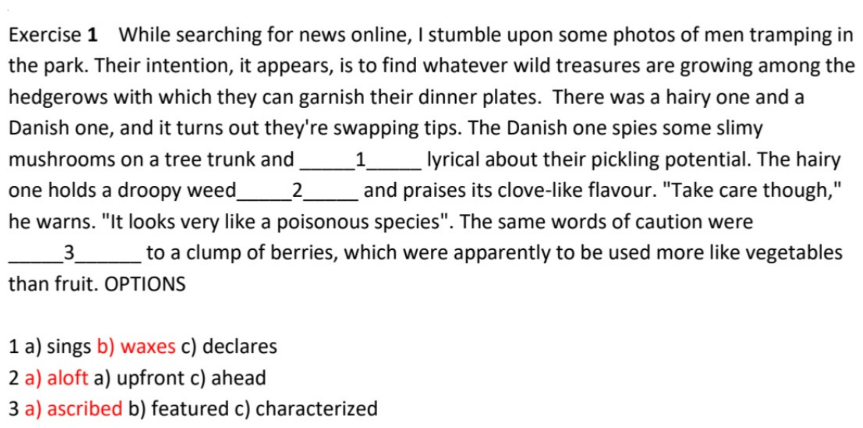
Figure B.6: Payment based on the median performance: Number of team members by gender



Note: P-value of two-sample Wilcoxon rank-sum (Mann-Whitney) test is 0.0436.

C Appendix: Screenshots

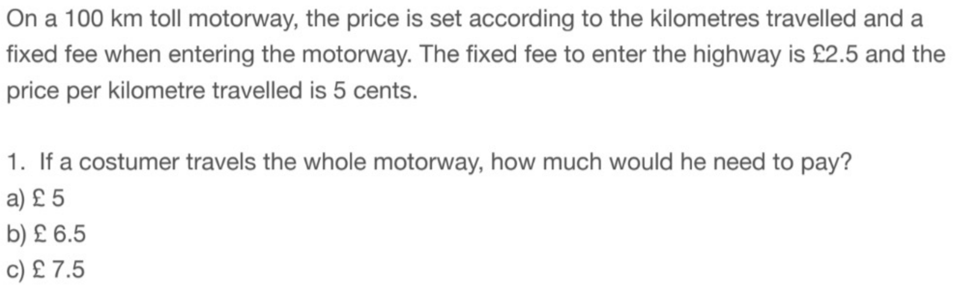
Figure C1: Example of Text task



Exercise **1** While searching for news online, I stumble upon some photos of men tramping in the park. Their intention, it appears, is to find whatever wild treasures are growing among the hedgerows with which they can garnish their dinner plates. There was a hairy one and a Danish one, and it turns out they're swapping tips. The Danish one spies some slimy mushrooms on a tree trunk and _____1_____ lyrical about their pickling potential. The hairy one holds a droopy weed_____2_____ and praises its clove-like flavour. "Take care though," he warns. "It looks very like a poisonous species". The same words of caution were _____3______ to a clump of berries, which were apparently to be used more like vegetables than fruit. OPTIONS

1 a) sings b) waxes c) declares
2 a) aloft a) upfront c) ahead
3 a) ascribed b) featured c) characterized

Figure C2: Example of Math task

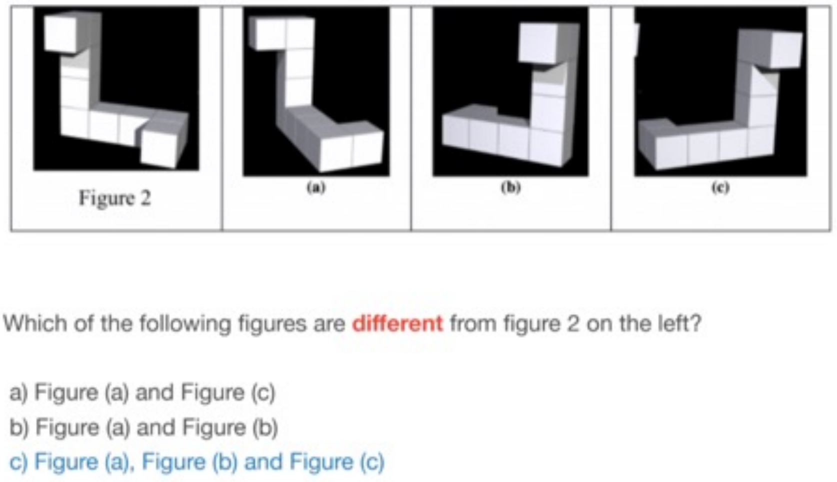


On a 100 km toll motorway, the price is set according to the kilometres travelled and a fixed fee when entering the motorway. The fixed fee to enter the highway is £2.5 and the price per kilometre travelled is 5 cents.

1. If a costumer travels the whole motorway, how much would he need to pay?

a) £ 5
b) £ 6.5
c) £ 7.5

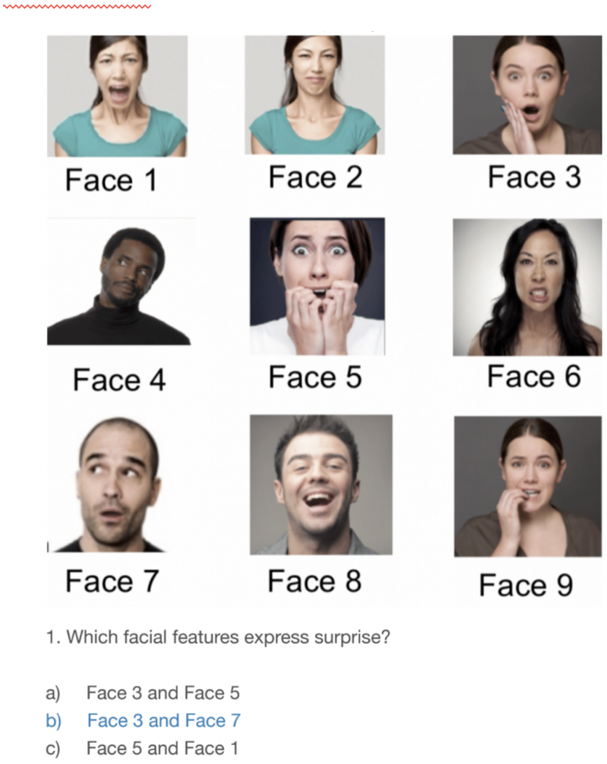
Figure C3: Example of Rotation task



Which of the following figures are **different** from figure 2 on the left?

a) Figure (a) and Figure (c)
b) Figure (a) and Figure (b)
c) Figure (a), Figure (b) and Figure (c)

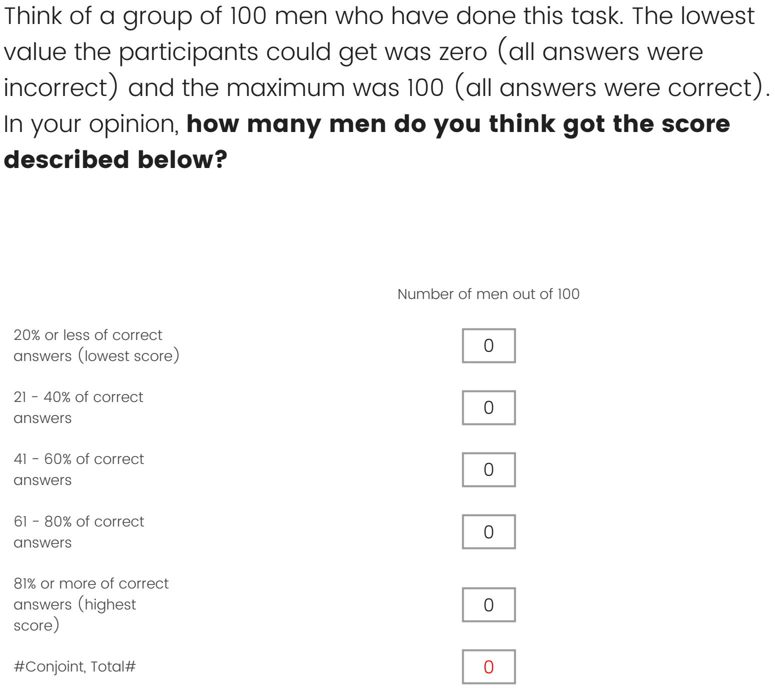
Figure C4: Example of Emotion task



1. Which facial features express surprise?

a) Face 3 and Face 5
b) Face 3 and Face 7
c) Face 5 and Face 1

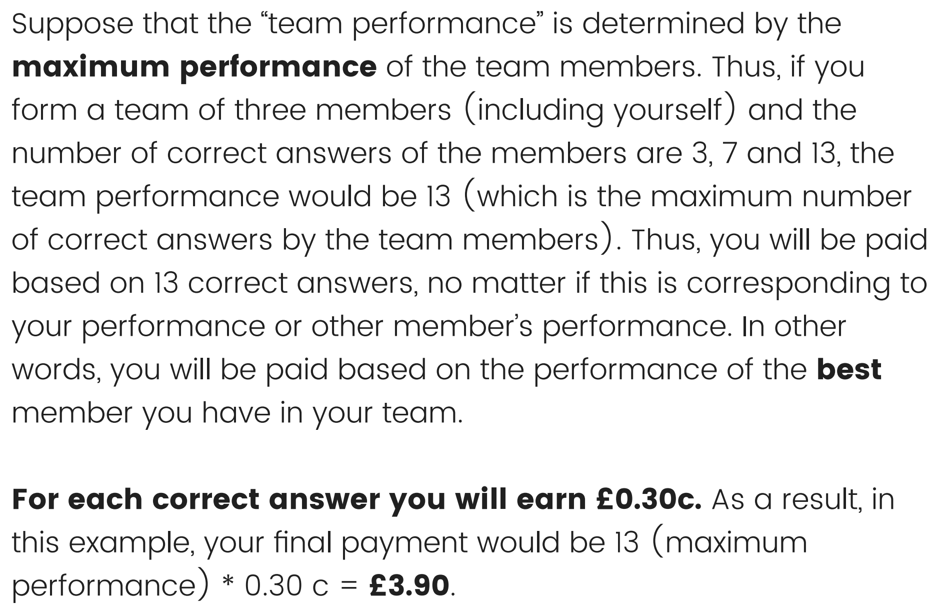
Figure C5: Example of Elicitation of Beliefs



Think of a group of 100 men who have done this task. The lowest value the participants could get was zero (all answers were incorrect) and the maximum was 100 (all answers were correct). In your opinion, **how many men do you think got the score described below?**

	Number of men out of 100
20% or less of correct answers (lowest score)	0
21 - 40% of correct answers	0
41 - 60% of correct answers	0
61 - 80% of correct answers	0
81% or more of correct answers (highest score)	0
#Conjoint, Total#	0

Figure C6: Team formation experiment – Example of Instructions for Maximum payment



Suppose that the "team performance" is determined by the **maximum performance** of the team members. Thus, if you form a team of three members (including yourself) and the number of correct answers of the members are 3, 7 and 13, the team performance would be 13 (which is the maximum number of correct answers by the team members). Thus, you will be paid based on 13 correct answers, no matter if this is corresponding to your performance or other member's performance. In other words, you will be paid based on the performance of the **best** member you have in your team.

For each correct answer you will earn £0.30c. As a result, in this example, your final payment would be 13 (maximum performance) * 0.30 c = **£3.90**.