Betting on Diversity – Occupational Segregation and Gender Stereotypes*

Urs Fischbacher† Dorothea Kübler‡ Robert Stüber§

Januar 11, 2021

Abstract

Many occupations and industries are highly segregated with respect to gender. This segregation could be due to perceived job-specific productivity differences between men and women. It could also result from the belief that single-gender teams perform better. We investigate the two explanations in a lab experiment with students and in an online experiment with personnel managers. The subjects bet on the productivity of teams of different gender compositions in tasks that differ with respect to gender stereotypes. We obtain similar results in both samples. Women are picked more often for the stereotypically female task and men more often for the stereotypically male task. Subjects do not believe that homogeneous teams perform better but bet more on diverse teams, especially in the task with complementarities. Elicited expectations about the bets of others reveal that subjects expect the effect of the gender stereotypes of tasks but underestimate others' bets on diversity.

JEL Classification:*C91; D9; J16*

Keywords: Gender segregation, hiring decisions, teams, discrimination, stereotypes

*We would like to thank participants of the conference "Gender Research in Labor and Experimental Economics" in Nürnberg 2018, the ESA World Meeting 2018 in Berlin, the THEEM Workshop 2018, the BBE Workshop, the VfS Annual Conference 2019, the Gender Gaps Conference 2021, and the ESA 2021 Global Around-the-Clock Virtual Meeting. We are also grateful to Andrea Ichino, Kai Barron, Simone Haeckl, Moritz Janas, Mara Rebaudo, Müge Süer, Christian Thöni, and Arthur Schramm for helpful comments on the study as well as to Jennifer Rontganger and Roberta Goettler for copy editing. We gratefully acknowledge financial support from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through the Excellence Cluster EXC 2035/1 "The Politics of Inequality" (Urs Fischbacher), the CRC TRR 190 "Rationality and Competition" (Dorothea Kübler), the Excellence Cluster EXC 2055 "Contestations of the Liberal Script" (Dorothea Kübler), and from Tamkeen under NYU Abu Dhabi Research Institute Award CG005 (Robert Stüber). We obtained ethical approval for the second part of the study with personnel managers by the WZB Research Ethics Committee (No. 2021/2/114). The second part was also preregistered in the AEA RCT Registry with ID AEARCTR-0007456 (see https://doi.org/10.1257/rct.7456-1.1).

†University of Konstanz, Box D-131, 78457 Konstanz, Germany, and Thurgau Institute of Economics; e-mail: urs.fischbacher@uni-konstanz.de

‡WZB, Reichpietschufer 50, 10785 Berlin, Germany, and Technical University Berlin; e-mail: kuebler@wzb.eu

§Center for Behavioral Institutional Design, NYU Abu Dhabi, PO Box 129188, UAE; e-mail: robert.stueber@nyu.edu

1 Introduction

Gender segregation characterizes labor markets around the world (Bureau of Labor Statistics, 2020). Blau and Kahn (2017) document both vertical segregation with a higher percentage of women in low-paying jobs and horizontal segregation between different occupations and industries.[1] In a World Bank report, Das and Kotikula (2019) advocate the reduction of gender segregation to strengthen female labor market participation and to eliminate the misallocation of talent, given the increased human capital of women. Vertical gender segregation is seen as costly because diversity in higher-level management has been linked to innovation and growth.[2]

Understanding what causes the widespread gender segregation in labor markets is important. Two observations speak in favor of an equilibrium phenomenon, which means that forces are present that stabilize or even increase the dominance of one gender in a profession. First, the jobs that are male or female dominated can vary considerably across countries.[3] Second, a correlation exists between male or female dominance in a profession and discrimination of the other gender (e.g., Riach and Rich, 2006). Whereas the first observation could be driven by the demand as well as the supply side, the second observation is clear evidence that employers play an important role in segregation.[4]

In this study, we investigate the role of performance expectations of employers in explaining gender segregation. Hiring decisions that reinforce gender roles can be driven by perceived job-specific productivity differences, such as men being better at some tasks and women better at others. An alternative explanation is that employers expect homogeneous teams to perform better and therefore hire workers from the dominant gender. With the help of two experiments, we disentangle the two explanations. We create counterfactual situations, for example stereotypically male jobs performed by women, and measure beliefs about the productivity of teams. Our experiment acknowledges the effect of a possibly unbalanced status quo, because segregation is widespread in many labor markets.

[1] About 50% of men work in occupations where less than 20% of the employees are women, and about 50% of women work in occupations where less than 30% of the employees are men (Cohen, 2013).

[2] Gender-balanced occupations and industries may provide other, less direct, benefits. For instance, increasing the share of male kindergarten teachers may be beneficial to give children the opportunity to interact equally with women and men. See Bayer and Rouse (2016) for a general discussion of the benefits of diversity of academic economists.

[3] For instance, women are underrepresented in the medical profession in the US, whereas they constitute the majority of practicing physicians in Russia and Eastern Europe (Ramakrishnan et al., 2014).

[4] On the supply side, policies have been suggested to target educational choices, to create a workplace culture that attracts both men and women, and to provide role models. However, a more equal distribution of men and women across occupations might be an important pull factor: more women in certain male occupations might inspire girls to aspire to these positions. Moreover, the importance of demand forces for gender convergence in recent decades has been emphasized (Olivetti and Petrongolo, 2016).

In the experiment, gender-homogeneous teams (that consist of four men or four women) work on a task. Then, we replace one worker with a new subject who is either a man or a woman. Thus, the team either becomes more diverse with a ratio of 3 to 1, or remains homogeneous. We elicit beliefs about the performance by letting the subjects bet on the performance of the newly formed teams.[5] We employ three tasks: assembling a bookshelf as a stereotypically male task, solving a memory game as a stereotypically female task, and writing down chains of movies and actors, which we designed with the objective of having women and men provide complementary inputs.[6]

Hiring decisions are likely to be influenced by productivity considerations, but they can also be affected by expectations regarding other people's beliefs about productivity differences. For example, employers might be reluctant to increase the diversity of their workforce or hire women for stereotypically male tasks even if they do not assume adverse productivity effects. They may simply act upon the assumption that their clients, superiors, employees, or stakeholders expect adverse effects, which, in turn, can be detrimental for the success of their business. For this reason, we also obtain incentivized measures of participants' beliefs about others' choices.

We conduct two versions of the experiment. The first version is a laboratory experiment with students at two German universities. However, students may have different views about hiring decisions than those who are actually responsible for such decisions. In particular, university students may value diversity more than managers. We therefore conduct an online experiment with personnel and general managers. They make hiring decisions, and their beliefs about the productivity of teams are important for explaining gender-segregated labor markets.

Our findings hold in both samples: gender stereotypes regarding the tasks are important. Women are expected to perform better in teams in the memory game, and men in assembling the shelves. At the same time, participants reveal a belief in the superiority of gender-diverse teams by choosing a woman more often in the initially all-male and a man more often in the initially all-female teams. The belief that diverse teams perform better is most pronounced for the network task, the task for which complementarities are im-

[5] Our design with a betting task allows us to focus only on beliefs about productivity and to exclude the possible influence of other factors such as discrimination based on taste, beliefs about the decisions of others, or social norms. For instance, when making hiring decisions in a lab experiment, subjects possibly trade off a taste for hiring one gender over the other with productivity considerations. Similarly, if hiring decisions are observed by others, social norms can come into play. By framing our experiment as a performance bet, beliefs about the bets of others do not influence choices. We focus on *gender* diversity but want to emphasize other forms of diversity including non-binary genders as valuable future research topics.

[6] Employing these different tasks allows us to test whether productivity estimates respond to stereotypes and whether there is an interaction between the job stereotypes and the desire for homogeneous or diverse teams. We use additional incentivized belief elicitations to investigate the stereotypes.

portant. Both the preference to follow the task stereotype and the preference for diversity are reflected in the participants' beliefs about others' choices.[7] However, whereas participants fully anticipate the effect of the task stereotype, they substantially underestimate the preference for diversity. This finding is particularly pronounced among the managers who, on average, expect others to believe that gender-homogeneous teams perform better. Finally, we observe an own-gender bias whereby male subjects more often bet on adding a male participant, whereas women more often believe that adding a woman will increase productivity.

The results indicate that the gender segregation observed in many occupations is not caused by beliefs in the higher productivity of homogeneous teams, because the subjects predict gender-diverse teams will perform better than homogeneous teams. Rather, the findings are consistent with segregation resulting from gender stereotypes regarding the tasks. These results, together with the finding that the preference for diverse teams is underestimated, can explain why gender imbalances in occupations are sustained. They imply that policies to reduce gender segregation may fail if they do not successfully alter stereotypes. Our results also emphasize that beliefs about team compositions, next to beliefs about productivity difference across tasks, have an impact on productivity assessments. This finding is important for a large literature on gender differences in the labor market that has largely ignored the highly unequal gender ratio in many occupations.

Beliefs and gender stereotypes take center stage in recent research on gender disparities in economic outcomes. Gender stereotypes can influence the gender gap in performance, as documented by the literature on stereotype threat (Spencer et al., 1999). They can also explain hiring decisions that disadvantage women (Reuben et al., 2014) and influence ability beliefs (Bordalo et al., 2019) and academic performance even if held by others (Carlana, 2019).[8] Coffman et al. (2021a), Bohren et al. (2019), and Bohren et al. (2020) provide evidence of discrimination that is based on biased beliefs rather than taste or accurate beliefs.[9] Barron et al. (2020) show that gender discrimination based on biased beliefs may be explicit or implicit. Although these studies provide important insights regarding the role of stereotypes and distorted beliefs for gender discrimination, they do not consider existing gender imbalances in occupations and the possibility that performance expectations depend on the gender composition of work teams.

Discrimination in gender-stereotypical occupations against the underrepresented gen-

[7] For convenience, we refer to the fact that individuals bet on the team for which an individual of the non-dominant gender is added as a "preference for diversity" and so on.

[8] Bordalo et al. (2016) present a model of stereotypes according to which stereotypes are rooted in true differences between teams but cause belief distortions by exaggerating small differences. Haeckl and Kartal (2021) study stereotypes and perseverance.

[9] Coffman et al. (2021a) also document in-group bias with respect to own gender.

der has been found in correspondence tests (Riach and Rich, 1987) and vignette studies (Kübler et al., 2018).[10] Our findings add to this literature by indicating that discrimination is most likely not driven by a preference for homogeneous work teams, but rather by the gender stereotypes of the tasks.

We use bets on the outcomes of team work to elicit beliefs regarding gender and productivity. Thus, we do not measure performance differences but focus on the beliefs about such differences. By contrast, differences in the performance of mixed and single-sex teams have been studied in the literature, both in the field (Adams and Ferreira, 2009, Apesteguia et al., 2012, and Hoogendoorn et al., 2013) and in the lab (Ivanova-Stenzel and Kübler, 2011). These studies report mixed results.[11] Moreover, Sarsons (2017), Isaksson (2018), Coffman et al. (2021b), and Sarsons et al. (2021) show that women claim and receive less credit for their joint work with men and that stereotypes can influence the degree to which team members of different genders receive recognition for the work. Whereas the focus of these studies is on differences in credit that women and men get for past work in gender-mixed teams, our study asks whether individuals expect differences in team productivity between homogeneous teams that remain homogeneous or become diverse.[12]

By analyzing expectations about performance assessments, our study relates to the findings of Dustan et al. (2021) who analyze beliefs about gender differences in individual performance. Dustan et al. (2021) document a difference in second-order beliefs between men and women in a scenario where they hold similar first-order beliefs. By contrast, we analyze the accuracy of second-order beliefs regarding team performance when task stereotype and team composition vary.

Finally, our study relates to laboratory experiments with different subject pools (see Gneezy et al., 2009, for a study in the context of gender inequality, and Fréchette, 2015 and 2016, for general overviews). Our finding that students and personnel managers hold similar beliefs about the performance of teams indicates the robustness of this feature and strengthens confidence in the external validity.

[10] Further vignette studies that test for discrimination in gender-stereotypical occupations are Cash et al. (1977), Muchinsky and Harris (1977), Sharp and Post (1980), and Glick et al. (1988). Correspondence tests on the topic include Riach and Rich (2006), Booth and Leigh (2010), Albert et al. (2011) and Carlsson (2011).

[11] Adams and Ferreira (2009) show that the average effect of gender diversity on firm performance is negative, and Apesteguia et al. (2012) document that three-women teams are outperformed by all other gender combinations, whereas Hoogendoorn et al. (2013) find that heterogeneous teams perform better than male-dominated teams. Ivanova-Stenzel and Kübler (2011) find that gender diversity increases the gender performance gap with team pay but lowers the gap with team competition.

[12] Our study therefore also relates to a recent literature on the productivity of teams (Gächter et al., 2019 and Weidmann and Deming, 2021). Moreover, the study connects to a large literature on homophily (e.g., McPherson et al., 2001 and Baccara and Yariv, 2013).

2 Experimental design, procedures, and samples

2.1 Treatments and sessions

Our primary goal is to understand how people assess the performance of all-male or all-female teams that either remain homogeneous or become diverse. We investigate the performance assessments of teams across tasks that are stereotypically male or female, or for which men and women are expected to provide complementary inputs. To this end, we run two experiments, each following a 2x3 factorial design. We measure the fraction of subjects who believe adding a man to a team is better than adding a woman for two initial team compositions (initially all-male or all-female) and three tasks. The subjects are asked to place bets on the relative performance of the teams. Moreover, they estimate the fraction of other subjects betting on each team.

The subjects in both experiments take on the role of evaluators in the "evaluator sessions" and are asked to predict which of the two teams will perform better. They are paid for correct predictions with respect to reference performances obtained in "team sessions." In the team sessions, individuals worked in teams on the tasks. These sessions were conducted only to incentivize the evaluators, and we therefore conducted only one session for each possible team composition. The instructions for the lab experiment with the students and the online experiment with the managers are provided in appendices C.1 and C.2.

We conducted the evaluator sessions with students in Berlin and Konstanz, Germany. For our main analyses, we combine the data from Konstanz and Berlin, but also investigate the datasets separately as a robustness check. The results from both subject pools are similar, although the composition of students differs (as described in Appendix B.2.2). Students may hold particular beliefs about the productivity of teams. In particular, the belief that gender-diverse teams perform better may be stronger for university students than for the general population. We therefore conduct an online experiment with personnel and general managers. Most of them regularly make hiring decisions, and their beliefs about the productivity of homogeneous or diverse teams can be responsible for gender-segregated labor markets.

The samples of 187 students and 443 managers are gender balanced, as shown in Table A.1 in the appendix. The students are, on average, 22 years old, whereas the average age of the managers is 46. The manager sample is more heterogeneous in terms of educational attainment, in that 60.50% obtained vocational training or lower as their highest educational degree, whereas around 40% obtained a university degree. The sample comprises higher- and lower-ranked managers, and 61.40% of them indicate that they had made hiring decisions in the past. As a robustness check, we provide additional results if we focus

on these individuals only.

We did not preregister the lab experiment, but the online experiment with personnel managers was preregistered (see https://doi.org/10.1257/rct.7456-1.1) because it may be harder to replicate. The entry at the AEA RCT Registry can be found in Appendix D. We never pool the data of the two experiments. We also obtained ethical approval for the online experiment (WZB Research Ethics Committee No. 2021/2/114).

2.2 Team sessions

2.2.1 General procedure

We conducted the team sessions with students in the lab. They consist of two parts. In part 1, a team of four men or four women performs a team task. The team's absolute performance determines the payoff of each team member. At the end of part 1, one of the four team members is randomly chosen. This subject receives his or her payment (show-up fee plus the payoff resulting from part 1) and is dismissed.

In part 2, a new person who was waiting in a separate room during part 1 is asked to enter the room and replace the dismissed team member. The new team member is either a man or a woman, which creates four combinations of initial team composition (all-male or all-female) and new member (male or female).[13] Then, the newly formed team performs a task similar to that in part 1.

The subjects participating in part 1 are informed about the procedures before the experiment starts, but they are not informed about the nature of the task performed in part 2. After part 2, the subjects receive their payment. For the three original team members and the new team member, the payment consists of the show-up fee, the payoff for part 1, and the payoff for part 2.

We employ three different tasks. They require the collaboration and interaction of the team members to ensure that expectations regarding the ability of teams to cooperate play a role for performance expectations. Furthermore, we chose tasks considered stereotypically male or female, or that require complementary inputs. The stereotypes can be driven by whether a man or a woman is expected to perform better in the task, or by whether a task is considered to be typically performed by this gender, and we provide empirical evidence for the stereotypes. Table 1 summarizes the main properties of the tasks.

[13] We ensure that the characteristics of the new team member (e.g., the gender) cannot influence the team performance in part 1, by inviting the subject to a different room. During the waiting time, the person reads the instructions the initial team members received for part 1.

Table 1. Tasks and predictions

Tasks	Gender stereotype	Complementarity
Shelves	Male	No
Memory	Female	No
Network	Neutral	Yes

Note: The table presents the predicted gender stereotypes and complementarities of men and women for the three tasks employed in the experiments.

2.2.2 Assembling shelves

In the shelves treatment, the four team members are asked to assemble two shelves of the same type as fast as possible. They can work simultaneously on both shelves and can make use of the manuals provided. Both shelves need to be fully built such that they are ready for use. The task is only accomplished if the shelves are correctly and completely built as assessed by the experimenter. Each subject's payoff function is $€\frac{100}{\text{min}}$, where min is the number of minutes needed to complete the task. If the shelves are not assembled completely after 40 minutes, the task is terminated and each subject receives €2.50. In part 2, the subjects have to assemble two shelves again. These shelves for part 2 are slightly different from the ones in part 1. The instructions and the payoff function are identical in the two parts.

We hypothesize that subjects perceive this task as stereotypically male. In Germany, only 2% of carpenters and 2% of construction workers are female. Similar proportions are observed in the UK and the US.[14] Assembling shelves is a physical task and requires spatial abilities. The stereotype that women perform worse in such tasks might therefore also arise because they are, on average, physically less strong than men and have lower spatial abilities (Voyer et al., 1995). Finally, men have been found to be better at assembling furniture (Wiking et al., 2016).

2.2.3 Memory game

Subjects in the memory treatment solve a memory game of 12x8 cards. The aim of the game is to find all pairs with the same picture with the fewest moves possible. A move consists of turning two cards that do not match. Each subject's payoff function is $€\frac{500}{\text{moves}}$. If the team

[14] See also https://www.theguardian.com/careers/careers-blog/2015/may/19/where-are-all-the-women-why-99-of-construction-site-workers-are-male and https://www.cnbc.com/2019/01/28/heres-what-its-like-to-be-a-woman-construction-worker.html.

needs less than five moves, each member receives €100. If the team needs more than 200 moves, the task is terminated and each subject receives €2.50. The team always decides jointly which two cards to reveal next. In the second part, the subjects solve a new memory game.

We assume subjects perceive this task as stereotypically female, especially because we remove any time pressure and base the payment solely on the number of steps needed (Shurchkov, 2012). Because the members decide jointly and every move reduces the team's payoff, the task requires a thorough and deliberate approach that is typical for many jobs that are considered stereotypically female, for instance administrative jobs. In line with this, Günther et al. (2010) and Iriberri and Rey-Biel (2017) used memory performance as a stereotypically female task.

2.2.4 Networks of films and actors

In this task, subjects are asked to write down a network that consists of screen productions (TV series, TV shows, and movies) as well as actresses and actors. Each screen production needs to be connected to an actress or actor who in turn needs to be connected to a screen production that needs to be connected to an actress or actor and so on. The team's goal is to build a chain as long as possible within the network in 10 minutes. When the 10 minutes are over, the experimenter checks all the connections. If a connection is wrong, for example, if an actress did not appear in the screen production, the network is broken. Each subject receives €0.40 for each correct connection in the network. Part 2 consists of creating another network in which connections between productions and actors that were part of the longest chain in part 1 cannot be used anymore.

We expect people to believe that men and women have complementary knowledge of screen productions and actors due to differences in their preferences. In particular, Wühr et al. (2017) provides evidence of actual gender differences in preferences for movie genres that are reflected in beliefs. Therefore, a diverse team may be expected to come up with more connections than an all-female or all-male team.

2.2.5 Procedures

We conducted the team sessions at the WZB-TU laboratory for experimental economics in Berlin. We ran one team session for each combination of team composition and task to incentivize the choices of the evaluators. We refrain from reporting the results, because we only observe one data point per treatment. The four shelves sessions lasted approximately 40 minutes, the memory sessions lasted 65 minutes, and the network sessions lasted 45

minutes. Subjects who participated in parts 1 and 2 earned, on average, €18.90 in the shelves sessions, €30.65 in the memory sessions, and €21 in the network sessions.

2.3 Evaluator sessions of laboratory experiment with students

2.3.1 Performance assessments

Our main research question is how individuals assess the productivity of teams that start out as all-male or all-female and remain homogeneous or become diverse. We measure individual beliefs regarding the productivity of these teams by letting subjects bet on whether a team of four people that was homogeneous in part 1 is more productive in part 2 if one of the four people is replaced by a person of the same or of the opposite gender. Subjects make a binary decision for both initial team compositions (all men or all women) and all three tasks. In particular, subjects choose between the sentence "The team with four women [men] will perform better in part 2 than the team with three women and one man [three men and one woman] in part 2" and "The team with three women and one man [three men and one woman] will perform better in part 2 than the team with four women [men] in part 2." For the sake of brevity, we use the expression "a subject chooses a man" for a subject betting on the team for which the new member is a man. We randomize the order of the tasks that subjects bet on.[15] For each of the six bets, subjects received €1.20 if the bet was correct. Table 2 provides an overview of the evaluator sessions.

Table 2. Overview of evaluator sessions

Part 1	Performance bets & quantitative performance assessments (*quantitative assessments only in lab experiment*)
Part 2	Elicitation of expectations about performance bets of others
Part 3	Bets regarding teams' stage 1 performance (*only in online experiment*)
Part 4	Questionnaires

Note: The table gives an overview of the experimental design in the evaluator sessions of the lab experiment and the online experiment.

We made sure the three tasks and the procedure were clear to the evaluators. Hence, subjects received the full instructions of all team sessions. We checked their understanding with a quiz that they had to answer correctly before they were allowed to proceed. Some

[15]The subjects make a decision for one initial team composition and task without knowing the other tasks. After having made a decision for all three tasks for the team composition, they make the three decisions for the other team composition.

of the team sessions were run after the evaluator sessions with students.[16] Thus, only after these evaluator sessions did the teams that were more productive become clear, and students in the role of evaluators were informed of the results only after all sessions had taken place. We thereby avoid information spillovers between evaluators in different sessions.[17] This procedure also emphasizes that the evaluators bet on performance differences, rather than making normative statements about whether a man or woman should be hired. In addition to incentivizing the elicitation of performance beliefs, this design reduces potential confounds due to social desirability.

After each bet, we elicited a measure of the size of the perceived difference between teams. We asked the evaluators to estimate the performance difference between the diverse and the homogeneous team. The subjects could choose between the categories "twice as good or more," "between twice as good and 50% better," "between 50% better and only slightly better," and "equally good." We did not incentivize this estimation.

2.3.2 Expectations about performance assessments

Hiring decisions can also be influenced by the expectations of employers regarding their clients' or superiors' views. We therefore elicit the subjects' expectations about performance assessments of others (second-order beliefs) by asking them to bet on the choices of others in the same session. The students are asked how many of 11 randomly chosen subjects in the session bet that the team to which a man was added performed better. They again receive €1.20 for every correct answer of the six team-task combinations.

For most analyses of the expectations about performance assessments, the dependent variable is the belief about how many out of 11 others bet that the team with a man performed better. An alternative approach is to consider a binary variable for the belief about the choices of the majority, which allows for additional comparisons between the binary choice and the (binarized) beliefs, because both are then measured on the same nominal scale. In some analyses, we also use this binarized belief measure.

2.3.3 Procedures

We conducted half of the evaluator sessions at the Lakelab at the University of Konstanz and half at the WZB-TU lab in Berlin. The sessions took place in the week from January 9 to January 16, 2018. We invited an equal number of men and women to each session.

[16] We ran some team sessions before the evaluator sessions to be sure the tasks that we proposed worked.

[17] For the online experiment with managers, running part of the evaluator sessions after the team sessions was not necessary, because information spillovers are unlikely.

The experiment was programmed using z-Tree (Fischbacher, 2007), and participants were invited using ORSEE (Greiner, 2015). Instructions were provided on-screen. After eliciting performance assessments and the expectations about the performance assessments of others, subjects completed a brief questionnaire including their gender, age, nationality, field of study, and semester.

Overall, 187 subjects (91 subjects in Konstanz, 96 in Berlin) participated in these sessions.[18] The sessions lasted around 40 minutes, and the average payment was €12.64; payments ranged from €8.20 to €21. Because some of the team sessions took place after the evaluator sessions, evaluators received the variable part of their payment three weeks after their session.[19]

2.4 Evaluator session of online experiment with managers

2.4.1 Performance assessments

We also ran an online experiment with personnel and general managers. To measure how the managers assess the productivity of the different teams, we ask them to make exactly the same six choices as the students (one for the team initially consisting of four men or four women, respectively, for all of the three tasks). Taking into account that managers may be more time-constrained than students and that the experiment was run online, the instructions for the managers were shorter and simpler than the students' instructions. Although the description of the tasks was the same as for students, the decision screen in the online experiment included images to illustrate the setup, whereas we relied only on text for the lab experiment (see Figure A.1 and Figure A.2 in Appendix A.1). We also did not include a quiz to check the managers' understanding. Like the students, the managers received €1.20 for each correct bet.

We did not elicit the unincentivized, quantitative measure of the size of the perceived difference between teams. Instead, we asked the managers to bet on whether the team consisting of four men or the team consisting of four women performed better in *stage 1* for each of the three tasks. This allows us to check whether assembling shelves is in fact considered a male task whereas solving memory games is a female task. The managers received €0.60 for each correct bet.

[18] We planned to conduct the experiment with 200 subjects. Because of no-shows, we ended up with 187.

[19] The subjects could choose between picking up the money in person, receiving an Amazon voucher, or receiving the money via a bank transfer. All subjects received a fixed payment of €7 (€5 show-up fee and €2 for completing the experiment) directly at the end of the session.

2.4.2 Expectations about performance assessments

After the managers' bets about the teams' stage 2 performances and before their bets about the teams' stage 1 performances, we elicited their expectations about others' performance assessments. As in the lab experiment, they were asked to assess how many of 11 other subjects bet on the team to which a man was added.[20] They received €0.60 for every correct answer.

2.4.3 Procedures

Based on our findings in the student sample, we conducted a power analysis (see the preregistration provided in Appendix D) indicating a sample of about 400 observations provides sufficient statistical power to test our main hypotheses. We collaborated with the data-collecting agency Respondi that approached personnel managers and general managers.[21] As for the lab experiment, we obtained a gender-balanced sample. Data collection took place between June 21 and June 23, 2021. The experiment was programmed using oTree (Chen et al., 2016). At the end of the experiment, we elicited participants' gender, age, occupation, and educational attainment. Because we are interested in individuals who make hiring decisions, we also elicited the degree to which the participants self-reported regularly making hiring decisions. Finally, because we conducted the experiment online, we also asked the managers to rate how much attention they paid to the questions when participating in the experiment.[22]

The median participant took about three minutes to complete the experiment for the session without belief elicitation and five minutes for the session with belief elicitation. The participants received a fixed payment of €0.25 for completing the experiment, and the average variable payoff was €5.34 (ranging from €1.20 to €9.60).

[20] To be able to incentivize these choices in the online experiment, we first needed to obtain some bets as a reference. We therefore obtained the first approximately 20 observations without belief elicitation. As preregistered, we use these observations in the analysis.

[21] Originally, we had planned and preregistered to conduct the experiment using email addresses of individuals posting a job offer on an online job platform. However, the response rate was very low. To be able to reach our targeted sample size, we then decided to collaborate with Respondi and preregistered this change before conducting the experiment.

[22] Participants could choose between "very inattentive," "inattentive," "rather inattentive," "neutral," "rather attentive," "attentive," and "very inattentive."

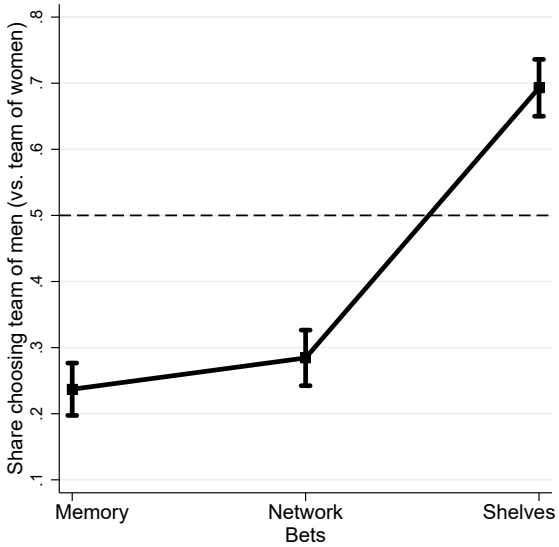
3 Results

3.1 Validation of task stereotypes: Expectations about stage 1 performance

We chose the three tasks, solving a memory game, building a network of actors and movies, and assembling shelves, to differ with respect to their gender stereotypes and the complementarity of female and male team members. To check whether the tasks are in fact gender stereotyped, we include the following measure in the online experiment: the participants were asked to bet on whether the all-male or all-female team performed better in stage 1 of the team sessions. We incentivize the choices of evaluators by comparing them with the average performance of the teams for every task.

As shown in Figure 1, most managers believe the all-female team performs better in the memory task and the all-male team in the shelves task (two-sided McNemar's test, $p < 0.001$). Hence, a substantial difference in gender stereotypes exists between the two tasks. We also observe that most managers believe the all-female team performs better than the all-male team in the network task, similar to the memory task. We use the network task mainly to study beliefs about diversity. The main experiment with bets on stage 2, which we report on below, reveals whether participants indeed believe diverse teams are relatively more productive in the network task than in the memory and the shelves tasks.

Figure 1. Stereotypes



Note: The figure shows the percentage of personnel managers expecting the all-male team to be more productive in stage 1 than the all-female team, along with 95%-confidence intervals.

3.2 Task stereotypes

Are task stereotypes responsible for gender segregation and the persistence of homogeneous work teams?[23] Panel (a) of Figure 2 shows that, pooling the choices for the two initial team compositions, only 36.36% of students bet on a man for the memory task whereas 67.91% bet on a man for the shelves task.[24] Indeed it is significant that more subjects believe that adding a man is better for the stereotypically male task than for the female task (two-sided McNemar's test, $p < 0.001$). The hypothesis that the gender stereotypes regarding the memory and shelves task drive the productivity beliefs of teams thus finds support in the data. In addition, the network task resembles the memory task in that subjects tend to bet on a woman: only 39.30% choose a man, which is significantly different from 50% (two-sided binomial test, $p < 0.001$).

We find similar results in the online experiment with managers. Significantly more managers believe that adding a man is better for assembling the shelves than for the memory game (see the gray line in Panel (a) of Figure 2; $p < 0.001$). About 36% of managers bet on a man for the memory task and 52.48% bet on a man for the shelves task, the latter percentage being smaller than in the student sample. Finally, the managers' and students'choices are similar in the network task.[25]

The observed bets are similar to the students' expectations about the choices of others, as shown in the right panel of Figure 2. The students believe, on average, that out of 11 randomly chosen participants, 5.24 subjects (47.64%) bet on a man for the memory task and 7.80 subjects (70.90%) bet on a man for the shelves task (black dotted line; two-sided t-test, $p < 0.001$).[26] Furthermore, the difference between choices and beliefs is statistically

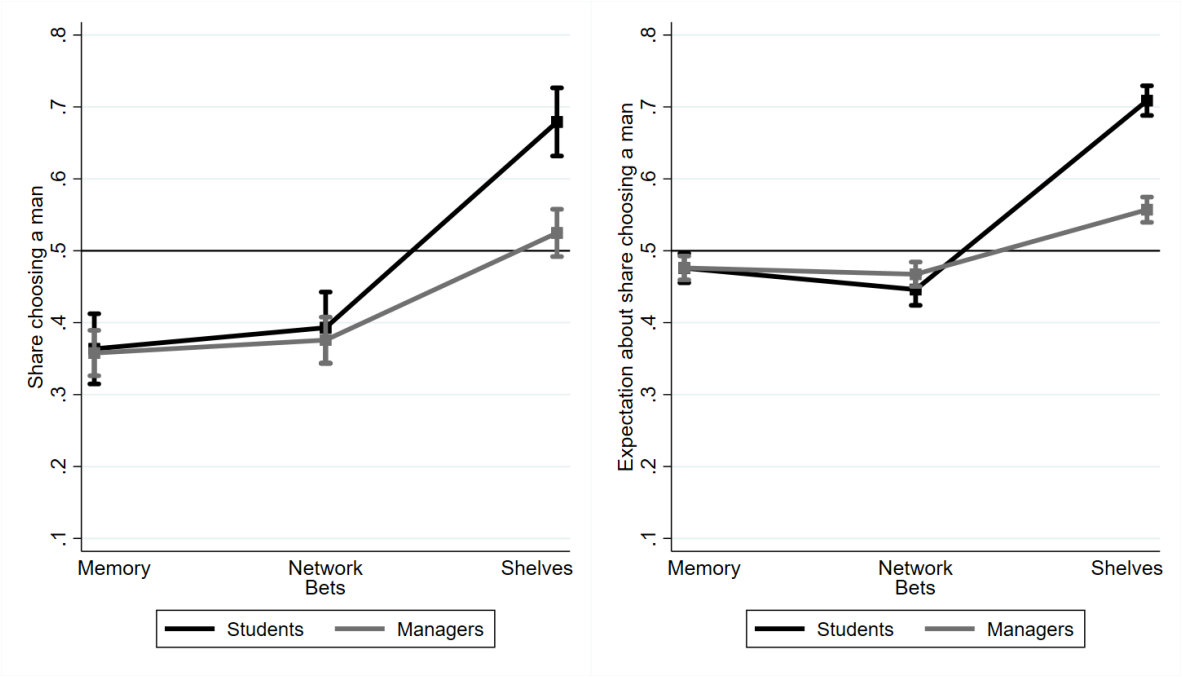
[23] Sections 3.2 to 3.4 present the results based on parametric and non-parametric tests. We sometimes pool several choices of one individual. For instance, we calculate the share betting on a man for the memory and the shelves task for which individuals make two decisions each, one for the initially all-female and one for the initially all-male team. All reported results are robust to using tests that take the dependence of choices into account. Moreover, the regressions that we report in section 3.5 are clustered at the individual level.

[24] The difference in expectations about male and female performance in the shelves and the memory tasks is also reflected in the subjects' quantitative assessments (see Figure A.3 in the appendix). The distribution of quantitative assessments is significantly shifted (Wilcoxon signed-rank test, $p < 0.001$, Kolmogorov-Smirnov test, $p < 0.001$). Because the distribution is discrete, we base the p-value for the latter test on a block permutation test and assign the quantitative assessments to the tasks at the level of the individual to take into account that one subject makes several choices.

[25] All these findings also hold if we compare choices between the memory and shelves task separately for the initially all-female and all-male team. In addition, for both samples, the difference between the shelves and memory task does not vary between the two initial team compositions.

[26] Figure A.4 in Appendix A.1 displays the shift in the distribution of beliefs about the choices of others induced by the task. The result also holds when we consider both initial team compositions separately (the difference is -2.83 for the initially male team and -2.29 for the initially female team) or if we consider binarized beliefs, that is, if the dependent variable is a dummy for whether an evaluator believes that the majority of other evaluators have chosen a man. The difference is 39 percentage points, whereas it is 32 percentage points based on the full belief information, and again highly statistically significant.

Figure 2. Bets and expectations by task and sample



(a) Bets (b) Expectations

Note: Panel (a) of the figure shows the share of students (black line) and managers (gray line) expecting the team to be more productive when a man is added pooled for both initial compositions of teams, along with 95%-confidence intervals. Panel (b) shows the average beliefs about this share for students (black line) and managers (gray line).

indistinguishable between the shelves and the memory task ($p = 0.125$).

Just as for the students, the managers' bets are consistent with their expectations about the choices of others, but the difference between memory and shelves task is attenuated, as shown by the gray line in Panel (b) of Figure 2. The difference between the beliefs in the two tasks is statistically significant in the pooled sample, but smaller.[27] Like the choices, the beliefs for the memory task (and the network task) are very similar between the student and manager sample, as can also be seen in Panel (b) of Figure 2. By contrast, the stereotype for the male shelves task is less pronounced in the manager sample.

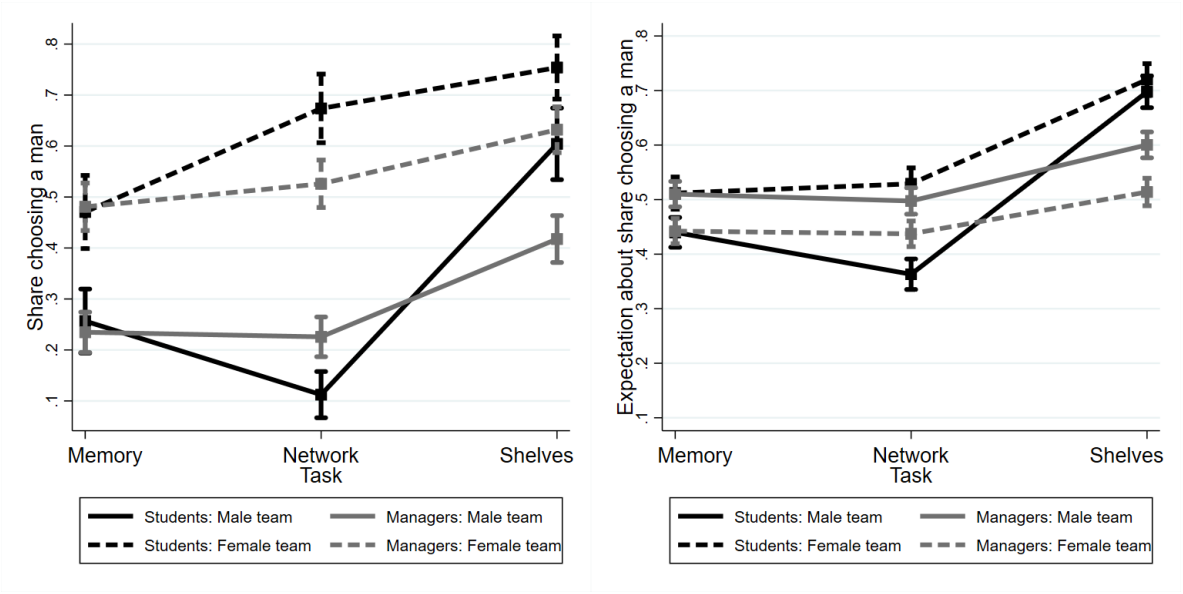
Result 1 [Task stereotypes]:
Most individuals select a man for the stereotypically male task (assembling shelves) and a woman for the stereotypically female task (memory) as well as for the network task. The difference in choices between the tasks is reflected in the beliefs about the choices of others.

[27] On average, the managers believe that out of 11 randomly chosen participants, 5.23 subjects (47.56%) bet on a man for the memory task and 6.13 subjects (55.71%) bet on a man for the shelves task. The beliefs for the memory and the network are very similar between the student and manager sample (see Panel (b) of Figure 2).

3.3 Team composition

We now turn to the question of whether people expect homogeneous teams to be more productive when they remain homogeneous than when they become more diverse. The left panel of Figure 3 plots the proportion of students and managers betting on a man for initially all-male and initially all-female teams by task. The black (gray) lines represent the students (managers), and the solid (dashed) lines are for the team that initially consisted of only women (men). For both samples and all three tasks, more students bet on a man if the initial team is female than if the initial team is male (two-sided McNemar's tests, $p < 0.001$). Averaging across the three tasks, we find that 65.42% of students and 62.68% of managers believe that mixed teams perform better than homogeneous teams. Both shares are significantly different from random choices ($p < 0.001$ and $p < 0.001$). Thus, subjects believe that diverse teams perform better.

Figure 3. Bets and expectations by task, team composition, and sample



(a) Bets

(b) Expectations

Note: Panel (a) shows the share of students (black lines) and managers (gray lines) betting on a man for the initially all-male team (solid lines) and the initially all-female team (dashed line) with 95%-confidence intervals. Panel (b) plots the average belief about the share betting on a man: black (gray) lines indicate students (managers) and solid (dashed) lines the beliefs for the initially all-male (all-female) team.

The preference for mixed teams varies across tasks. It is largest for the network task, as predicted. In particular, the proportion of students who believe adding a man leads to a higher performance than adding a woman if the team consists of three women rather than three men differs only by 21 and 15 percentage points for the memory and the shelves task, respectively. By contrast, this difference is more than twice these amounts for the network

task with 56% ($p < 0.001$ and $p < 0.001$).[28] In the manager sample, the difference is 30.02 percentage points for the network task, 24.61 percentage points for the memory task, and 21.45 percentage points for the shelves task, respectively. Therefore, the difference in the differences is smaller and not significant or marginally significant between the network and memory task ($p = 0.218$) and the network and the shelves task ($p < 0.057$), respectively.

Do subjects expect others to believe that diverse teams perform better? We find that the estimated proportion of students betting on a man is higher if the initial team consisted of women than when the initial team consisted of men for all three tasks (see Panel (b) of Figure 3).[29] Just as for the bets, the difference in the average belief about the number of other students choosing a man is significantly more pronounced for the network task ($p < 0.001$ and $p < 0.001$).[30] However, we find that students and managers underestimate the evaluators' beliefs in diverse teams on average (two-sided t-tests, $p < 0.001$), with managers even expecting that others believe homogeneous teams are more productive than diverse teams.

Students and managers are significantly better at predicting the effect of the task stereotype on choices than at predicting the desire for diversity (two-sided t-tests, $p < 0.001$ each). The students, for instance, correctly believe that others will choose a man if the team consists of women rather than of men in only 51.34% of cases. By contrast, students hold correct beliefs about the effect of the stereotype of a task on choices in 77.54% of cases. Students' mispredictions are driven solely by them underestimating the preference for diversity when the initial team consists of men: whereas the fraction of diversity choices does not vary between initially all-male and all-female teams (two-sided t-test, $p = 0.120$), the beliefs about the fraction of others betting on a diverse team are substantially lower if the initial team consists of men (two-sided t-test, $p < 0.001$).[31] For the managers, the mispredictions lead to them believing that others preferred homogeneous teams. This finding is independent of the initial team composition.

[28] The findings are corroborated when we compare the quantitative assessments for each of the three tasks between the initial team compositions (see Figure A.5). For all three tasks, we observe a shift in the distribution of assessments when comparing the initially male and female team in Wilcoxon signed-rank test and Kolmogorov-Smirnov tests. An ordered probit regression reveals that the shift in mean assessment is significantly larger for the network task.

[29] The difference is statistically significant for the memory and network task, but not statistically significant for the shelves task (two-sided t-test, $p = 0.275$).

[30] Considering the dummy for whether an evaluator believes the majority of others have chosen a man, we also find that the estimated proportion of evaluators betting on a man is higher if the initial team consisted of women than when the initial team consisted of men (two-sided t-test, $p < 0.001$).

[31] When we consider the binary variable for the belief about the choices of the majority, we find that students underestimate other students' preference for diversity if the initial team consists of men only (two-sided t-test, $p < 0.001$) but not if the initial team consists of women only ($p = 0.362$).

Result 2 [Team composition]:
Most evaluators predict that diverse teams perform better than homogeneous teams. The students significantly underestimate the degree to which others value diversity. The managers even expect that others predict homogeneous teams to perform better.

3.4 Gender of evaluator

Do women and men hold different beliefs about the productivity of teams? As can be taken from Panel (a) of Figure 4, for all six task-composition combinations, the share of male students betting on a man (gray lines) is higher than the fraction of female students betting on a man (black lines). Pooling across tasks and team compositions, we find that male students bet on a man in 53% of the cases, whereas female students do so in 43% of the cases. This difference of 10 percentage points is significant at the 1% level (two-sided Fisher's exact test, $p = 0.002$).[32] Equally, male managers (gray lines in Panel (c)) are significantly more likely to believe that the team to which a man was added was more productive (diff.: 6 pp; $p < 0.001$).[33]

With respect to the effect of the task stereotypes, the difference between the shelves and the memory task is similar for female (diff.: 32.10 pp) and male students (diff: 30.98 pp) and female (diff.: 16 pp) and male personnel managers (diff.: 18 pp). In addition, we document the belief that diverse teams are more productive for both female (66.14%) and male students (64.67%) and female (61.36%) and male managers (63.98%).

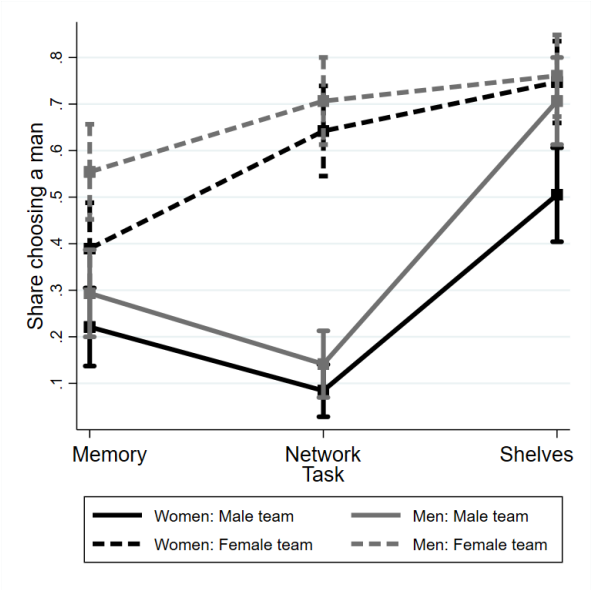
Regarding the beliefs about the bets of others, we do not observe a difference between male and female evaluators (Panel (b) and Panel (d) of Figure 4): for all 12 combinations of task, team composition, and sample, the expectations about performance assessments are similar, and the difference between male and female respondents is not significantly different from zero. [34] The effect of the task stereotypes is similarly reflected in both female and male students' and managers' expectations about the choices of others, whereas the belief that diverse teams are more productive is to a similar degree (not) reflected in the

[32] This also holds true if we consider the gender difference of evaluators separately for the three cases where the initial team consists of men and for the three cases where the initial team consists of women (see Figure 4, Panel (a)). For initially all-female teams, the difference is only marginally significant ($p = 0.054$). Considering the three tasks separately, the difference is significantly different from zero for the shelves task when the initial team consists of men only and for the memory task when the initial team consists of women.

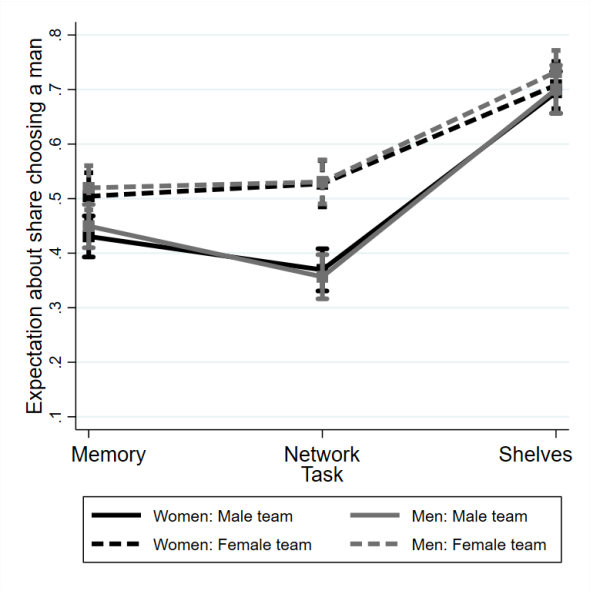
[33] The difference in choices between female and male evaluators is in line with an in-group bias regarding gender for individual performance assessments (Coffman et al., 2021a).

[34] Male students, on average, believe that 58.70% of subjects bet on a man, whereas female students believe that 55.44% of participants bet on a man($p = 0.278$). Male managers expect that others bet on a man in 50.48% of cases and female managers in 49.57% of cases ($p = 0.372$). The lab sessions were gender balanced, so for the lab experiment, students should have formed a belief about 11 others, of which half are women. The results are consistent with the managers also basing their beliefs on a gender-mixed group of others.

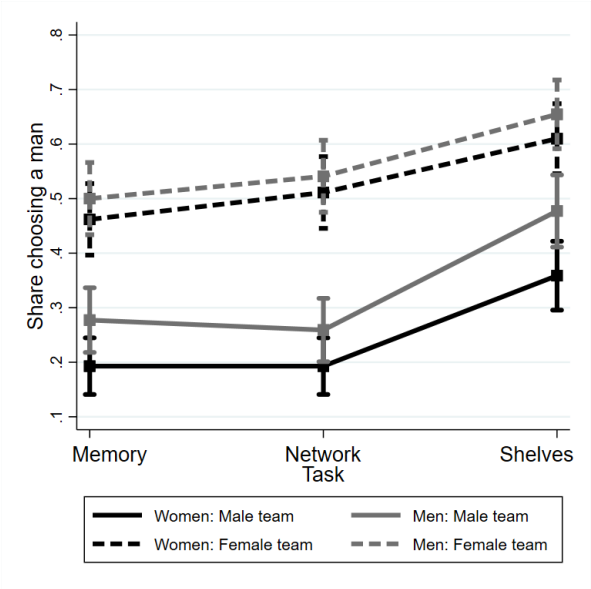
Figure 4. Bets and expectations by gender



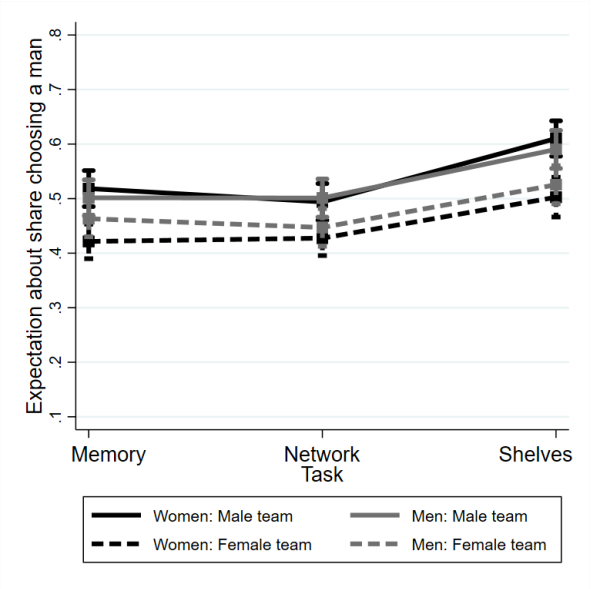
(a) Students: Bets



(b) Students: Expectations



(c) Managers: Bets



(d) Managers: Expectations

Note: Panel (a) shows the share of female (black lines) and male (gray lines) students expecting the team to be more productive when a man is added for initially all-male (solid lines) and initially all-female teams (dashed lines). Panel (c) shows the same for female and male managers. Panel (b) shows female students' (black lines) and male students' (gray lines) beliefs about the share of others expecting the team to be more productive when a man is added for initially all-male (solid lines) and initially all-female teams (dashed lines). Panel (d) presents the same data for managers.

expectations of men and women regarding the others' choices.

Result 3 [Gender of evaluator]:
Most evaluators display in-group bias regarding their own gender when betting on a man or a woman, but no gender difference exists in the expectations about other evaluators.

3.5 Robustness

3.5.1 Regression analysis

To investigate the robustness of the findings, we conduct a regression analysis detailed in Appendix B.1. There, we replicate the results from sections 3.2 to 3.4. We show that, for both students and managers, the results are robust to (i) controlling for the fact that the subjects make multiple decisions by clustering standard errors by subject, (ii) controlling for choice-order fixed effects (i.e., dummies for whether the choice for the given combination of task and initial team composition was made first, second, etc.) and session fixed effects, and (iii) controlling for individual-level variables. In both samples, we find no evidence that individual characteristics other than the gender of the subjects significantly correlate with the probability of betting on a man.

As part of the regression analysis, we also test for differences in the effect of the task stereotype and of the initial team composition between the student and the professional sample. Our results show that the effect of the team composition is similar for students and managers. The effect of changing the task stereotype (male vs. female) is significantly more pronounced for the student sample.

3.5.2 Further robustness checks

We engage in four further robustness checks. First, in Appendix B.2.1, we exploit the fact that we simultaneously observe several choices of each subject and classify evaluators as following task stereotypes or not and preferring homogeneous or diverse teams. We find that about two-thirds of the subjects make choices that are either consistent with the task stereotypes or with a desire for diverse teams, or both.

Second, for the lab experiment, we explore differences between our two subject pools, the pool in Berlin and Konstanz, which differ in their student composition. The analyses, detailed in Appendix B.2.2, reveal that the results are virtually identical for both subject pools.

Third, we investigate the robustness of our findings of the online experiment in Appendix B.2.3. We show that the results hold also for the restricted sample of personnel and general managers who regularly make hiring decisions, and of those managers who pass the attention check. We also exploit the fact that the manager sample is more heterogeneous than the student sample, and engage in an exploratory analyses of how different individual characteristics of this sample moderate the effect of team composition and task stereotype on productivity beliefs. We find little evidence that this heterogeneity is meaningful for beliefs about team performance. Rather, our findings are robust across different subsamples.

Fourth, we investigate the role of order effects by restricting the sample to subjects' first choices in Appendix B.2.4. We find that order effects seem to be non-negligible for the effect of the task stereotypes, in particular in the manager sample. We do not find any order effects for the effect of team diversity on productivity beliefs. In the appendix, we also discuss potential explanations for these findings.

4 Discussion

Gender stereotypes of tasks influence performance expectations. Participants believe that a team will be more productive if a man rather than a woman replaces an existing team member if the task is stereotypically male, and vice versa. Moreover, most subjects believe that mixed-gender teams show a higher performance than homogeneous teams. This is most pronounced for the task for which we predicted complementarities. In addition, we observe a preference of subjects to bet on workers of their own gender. Regarding expectations about others, we find that the effect of task stereotypes on other evaluators' choices is anticipated, whereas the subjects substantially underestimate the preference for diversity (the students) or even expect others to assign higher productivity to gender-homogeneous teams (the managers).

Our findings are relevant for a large literature analyzing the effect of (potentially biased) beliefs on gender discrimination. To analyze hiring decisions and performance beliefs, these studies typically consider different tasks with different gender stereotypes. By changing the stereotypes, this literature may have also implicitly changed beliefs about the workforce that typically performs these tasks. Studies therefore may have been also picking up the effect of the team composition.

The results suggest that gender segregation in labor markets is the result of task-related gender stereotypes rather than of beliefs about the superiority of homogeneous teams. Policies that attempt to reduce gender segregation need to alter these stereotypes. Pro-

viding role models may be particularly helpful.[35] Quotas will only permanently increase the female share in male-dominated occupations if they change the stereotypes.

With respect to issues that are highly topical, such as gender, employers may prefer to stick to what others think is the right thing to do. In this sense, second-order beliefs can be a source of and contribute to the strength of performance expectations. Our findings suggest that task stereotypes may dominate the desire for diversity, implying that gender imbalances can be persistent. The degree to which first- and second-order beliefs are relevant for hiring decisions is an interesting question for future work.

The fact that the results from the lab experiment with students and the online experiment with managers are consistent speaks to the generalizability and replicability of our findings.[36] Together with the fact that the betting tasks are incentivized and framed as a prediction task and not as a hiring decision, we are confident that the subjects' answers are not driven by social desirability.[37] Yet, some caveats apply. First, gender is salient for the evaluators because we mention it explicitly. Therefore, we might overstate the role of gender for the performance estimates. Second, although we intentionally chose three tasks that resemble real-world occupations, investigating other tasks that are even closer to real jobs seems worthwhile. Finally, we only study the extreme cases of initially all-male or all-female teams. Many open questions remain, such as questions relating to performance beliefs when teams are mixed at the outset or when they are larger.

[35] See, for instance, https://inspiring-girls.com/.

[36] In addition, our findings in the lab experiment are virtually identical for Konstanz and Berlin, two cities that are different in various respects.

[37] For example the finding that most subjects bet on a man for the stereotypically male task is not in line with common views of what is politically correct. Note that in an experiment conducted with the same Berlin subject pool, though with different subjects, and in parallel to our experiment, Barron et al. (2020) document explicit discrimination against women; that is, subjects more often bet that a man will be more productive when choosing between an equally qualified man and woman. Hence, it is not the case that the subjects in the subject pool do not discriminate against women in lab experiments. In addition, we observe more diverse choices for the task that requires complementarities.

References

ADAMS, R. B. AND D. FERREIRA (2009): "Women in the boardroom and their impact on governance and performance," *Journal of Financial Economics*, 94, 291–309.

ALBERT, R., L. ESCOT, AND J. A. FERNÀNDEZ-CORNEJO (2011): "A field experiment to study sex and age discrimination in the Madrid labour market," *The International Journal of Human Resource Management*, 22, 351–375.

APESTEGUIA, J., G. AZMAT, AND N. IRIBERRI (2012): "The impact of gender composition on team performance and decision making: Evidence from the field," *Management Science*, 58, 78–93.

BACCARA, M. AND L. YARIV (2013): "Homophily in peer groups," *American Economic Journal: Microeconomics*, 5, 69–96.

BARRON, K., R. DITLMANN, S. GEHRIG, AND S. SCHWEIGHOFER-KODRITSCH (2020): "Explicit and implicit belief-based gender discrimination: A hiring experiment," WZB Discussion Paper SP II 2020-306.

BAYER, A. AND C. E. ROUSE (2016): "Diversity in the Economics Profession: A New Attack on an Old Problem," *Journal of Economic Perspectives*, 30, 221–42.

BLAU, F. D. AND L. M. KAHN (2017): "The Gender Wage Gap: Extent, Trends, and Explanations," *Journal of Economic Literature*, 55, 789–865.

BOHREN, J. A., K. HAGGAG, A. IMAS, AND D. G. POPE (2020): "Inaccurate statistical discrimination," Working paper.

BOHREN, J. A., A. IMAS, AND M. ROSENBERG (2019): "The dynamics of discrimination: Theory and evidence," *American Economic Review*, 109, 3395–3436.

BOOTH, A. AND A. LEIGH (2010): "Do employers discriminate by gender? A field experiment in female-dominated occupations," *Economics Letters*, 107, 236–238.

BORDALO, P., K. COFFMAN, N. GENNAIOLI, AND A. SHLEIFER (2016): "Stereotypes," *The Quarterly Journal of Economics*, 131, 1753–1794.

——— (2019): "Beliefs about gender," *American Economic Review*, 109, 739–73.

BUREAU OF LABOR STATISTICS (2020): Household data, annual averages 2019; Retrieved from https://www.bls.gov/cps/cpsaat11.htm.

CARLANA, M. (2019): "Implicit stereotypes: Evidence from teachers' gender bias," *The Quarterly Journal of Economics*, 134, 1163–1224.

CARLSSON, F., M. R. MØRKBAK, AND S. B. OLSEN (2012): "The first time is the hardest: A test of ordering effects in choice experiments," *Journal of Choice Modelling*, 5, 19–37.

CARLSSON, M. (2011): "Does hiring discrimination cause gender segregation in the Swedish labor market?" *Feminist Economics*, 17, 71–102.

CASH, T. F., B. GILLEN, AND D. S. BURNS (1977): "Sexism and beautyism in personnel consultant decision making." *Journal of Applied Psychology*, 62, 301.

CHEN, D. L., M. SCHONGER, AND C. WICKENS (2016): "oTree–An open-source platform for laboratory, online, and field experiments," *Journal of Behavioral and Experimental Finance*, 9, 88–97.

COFFMAN, K., C. L. EXLEY, AND M. NIEDERLE (2021a): "The Role of Beliefs in Driving Gender Discrimination," *Management Science*, 67, 3551–3569.

COFFMAN, K., C. B. FLIKKEMA, AND O. SHURCHKOV (2021b): "Gender stereotypes in deliberation and team decisions," *Games and Economic Behavior*, 129, 329–349.

COHEN, P. N. (2013): "The Persistence of Workplace Gender Segregation in the US," *Sociology Compass*, 7, 889–899.

DAS, S. AND A. KOTIKULA (2019): "Gender-based employement segregation: understanding causes and policy interventions," Jobs Working Paper 26, World Bank Group.

DUSTAN, A., K. KOUTOUT, AND G. LEO (2021): "Beliefs about Beliefs about Gender," Working paper.

FISCHBACHER, U. (2007): "z-Tree: Zurich toolbox for ready-made economic experiments," *Experimental Economics*, 10, 171–178.

FRÉCHETTE, G. R. (2015): "Laboratory Experiments: Professionals versus Students," in *The Methods of Modern Experimental Economics*, ed. by G. Fréchette and A. Schotter, Oxford University Press.

——— (2016): "Experimental Economics across Subject Populations," in *The Handbook of Experimental Economics, Volume 2*, ed. by J. H. Kagel and A. E. Roth, Princeton University Press, 435–480.

GÄCHTER, S., C. STARMER, AND F. TUFANO (2019): "The surprising capacity of the company you keep: revealing group cohesion as a powerful factor of team production," CeDex Discussion Paper 2019-16, CeDex.

GLICK, P., C. ZION, AND C. NELSON (1988): "What mediates sex discrimination in hiring decisions?" *Journal of Personality and Social Psychology*, 55, 178.

GNEEZY, U., K. L. LEONARD, AND J. A. LIST (2009): "Gender Differences in Competition: Evidence From a Matrilineal and a Patriarchal Society," *Econometrica*, 77, 1637–1664.

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

GÜNTHER, C., N. A. EKINCI, C. SCHWIEREN, AND M. STROBEL (2010): "Women can't jump?–An experiment on competitive attitudes and stereotype threat," *Journal of Economic Behavior & Organization*, 75, 395–401.

HAECKL, S. AND M. KARTAL (2021): "Does a stereotype benefit women in the labor market: An experiment on perseverance," Unpublished manuscript.

HOOGENDOORN, S., H. OOSTERBEEK, AND M. VAN PRAAG (2013): "The impact of gender diversity on the performance of business teams: Evidence from a field experiment," *Management Science*, 59, 1514–1528.

IRIBERRI, N. AND P. REY-BIEL (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.

ISAKSSON, S. (2018): "It Takes Two: Gender Differences in Group Work," Unpublished manuscript.

IVANOVA-STENZEL, R. AND D. KÜBLER (2011): "Gender differences in team work and team competition," *Journal of Economic Psychology*, 32, 797–808.

KÜBLER, D., J. SCHMID, AND R. STÜBER (2018): "Gender discrimination in hiring across occupations: a nationally-representative vignette study," *Labour Economics*, 55, 215–229.

MCPHERSON, M., L. SMITH-LOVIN, AND J. M. COOK (2001): "Birds of a feather: Homophily in social networks," *Annual review of sociology*, 27, 415–444.

MUCHINSKY, P. M. AND S. L. HARRIS (1977): "The effect of applicant sex and scholastic standing on the evaluation of job applicant resumes in sex-typed occupations," *Journal of Vocational Behavior*, 11, 95–108.

OLIVETTI, C. AND B. PETRONGOLO (2016): "The Evolution of Gender Gaps in Industrialized Countries," *Annual Review of Economics*, 8, 405–434.

RAMAKRISHNAN, A., D. SAMBUCO, AND R. JAGSI (2014): "Women's participation in the medical profession: insights from experiences in Japan, Scandinavia, Russia, and Eastern Europe," *Journal of Women's Health*, 23, 927–934.

REUBEN, E., P. SAPIENZA, AND L. ZINGALES (2014): "How stereotypes impair women's careers in science," *Proceedings of the National Academy of Sciences*, 111, 4403–4408.

RIACH, P. A. AND J. RICH (1987): "Testing for sexual discrimination in the labour market," *Australian Economic Papers*, 26, 165–178.

——— (2006): "An experimental investigation of sexual discrimination in hiring in the English labor market," *Advances in Economic Analysis & Policy*, 5.

SARSONS, H. (2017): "Recognition for group work: Gender differences in academia," *American Economic Review*, 107, 141–45.

SARSONS, H., K. GËRXHANI, E. REUBEN, AND A. SCHRAM (2021): "Gender differences in recognition for group work," *Journal of Political Economy*, 129, 101–147.

SHARP, C. AND R. POST (1980): "Evaluation of male and female applicants for sex-congruent and sex-incongruent jobs," *Sex Roles*, 6, 391–401.

SHURCHKOV, O. (2012): "Under pressure: Gender-differences in output quality and quantitiy under competition and time constraints," *Journal of the European Economic Association*, 10, 1189–1213.

SPENCER, S. J., C. M. STEELE, AND D. M. QUINN (1999): "Stereotype threat and women's math performance," *Journal of Experimental Social Psychology*, 35, 4–28.

THE SYDNEY MORNING HERALD (2008): "Women best at assembling IKEA furniture," October 10, 2021, retrieved from: https://www.smh.com.au/world/women-best-at-assembling-ikea-furniture-20081226-759i.html.

VOYER, D., S. VOYER, AND M. P. BRYDEN (1995): "Magnitude of sex differences in spatial abilities: a meta-analysis and consideration of critical variables." *Psychological Bulletin*, 117, 250.

WEIDMANN, B. AND D. J. DEMING (2021): "Team Players: How Social Skills Improve Team Performance," *Econometrica*, 89, 2637–2657.

WIKING, S., M. L. BRATTFJELL, E. E. IVERSEN, K. MALINOWSKA, R. L. MIKKELSEN, L. P. RØED, AND J. E. WESTGREN (2016): "Sex Differences in Furniture Assembly Performance: An Experimental Study," *Applied Cognitive Psychology*, 30, 226–233.

WÜHR, P., B. P. LANGE, AND S. SCHWARZ (2017): "Tears or fears? Comparing gender stereotypes about movie preferences to actual preferences," *Frontiers in Psychology*, 8, 428.

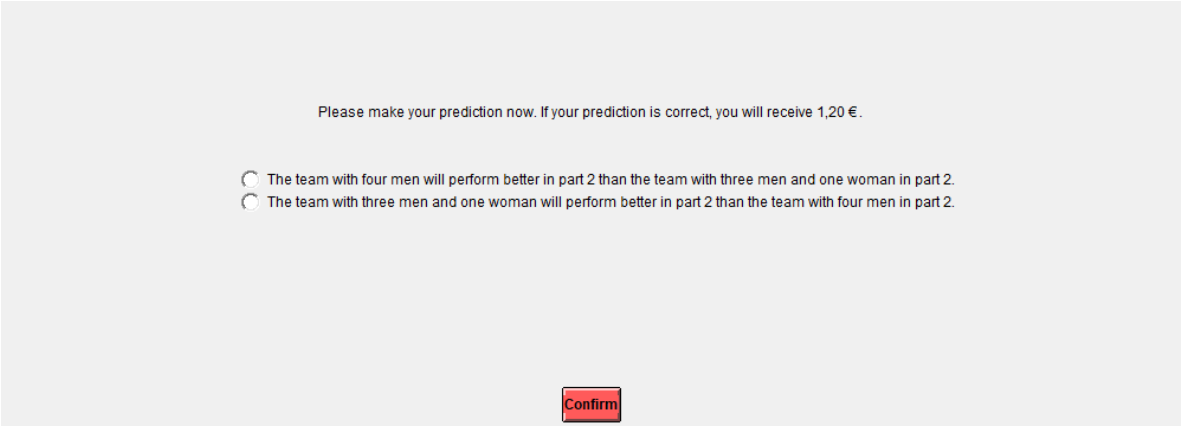
ZIZZO, D. J. (2010): "Experimenter demand effects in economic experiments," *Experimental Economics*, 13, 75–98.

APPENDIX

A Additional figures and tables

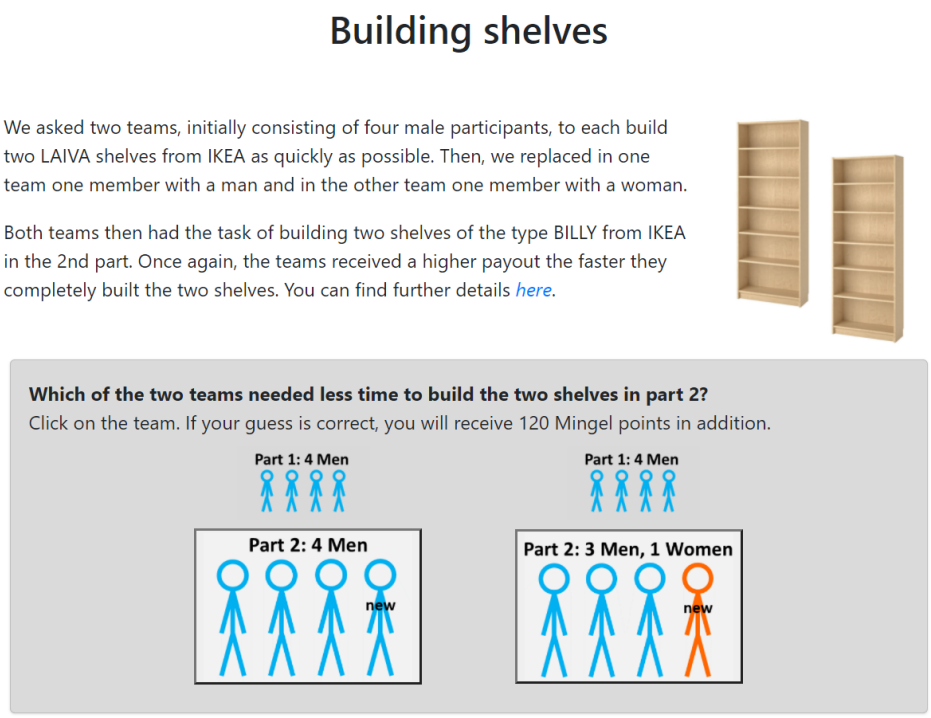
A.1 Additional figures

Figure A.1. Decision screen laboratory experiment



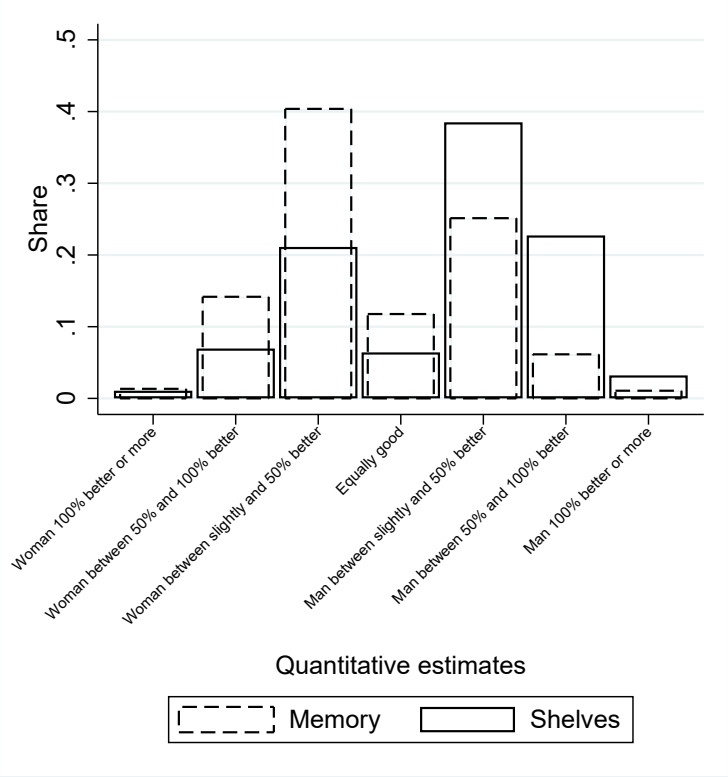
Note: The figure shows the decision screen in the laboratory experiment.

Figure A.2. Decision screen online experiment



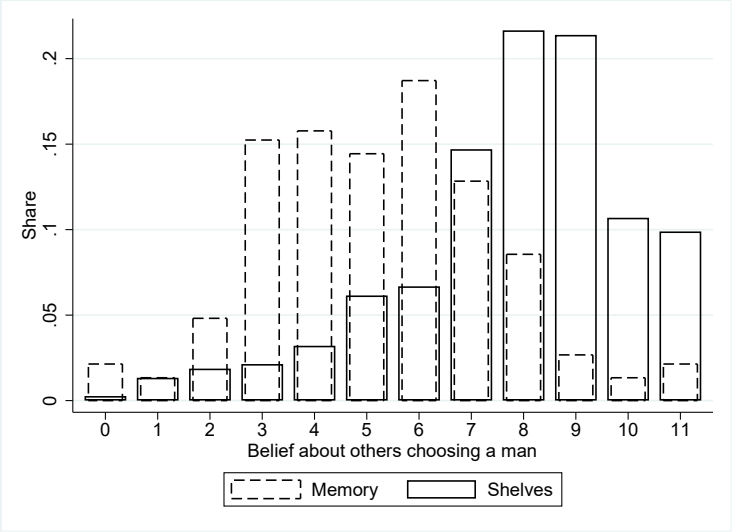
Note: The figure shows the decision screen in the online experiment.

Figure A.3. Quantitative estimates



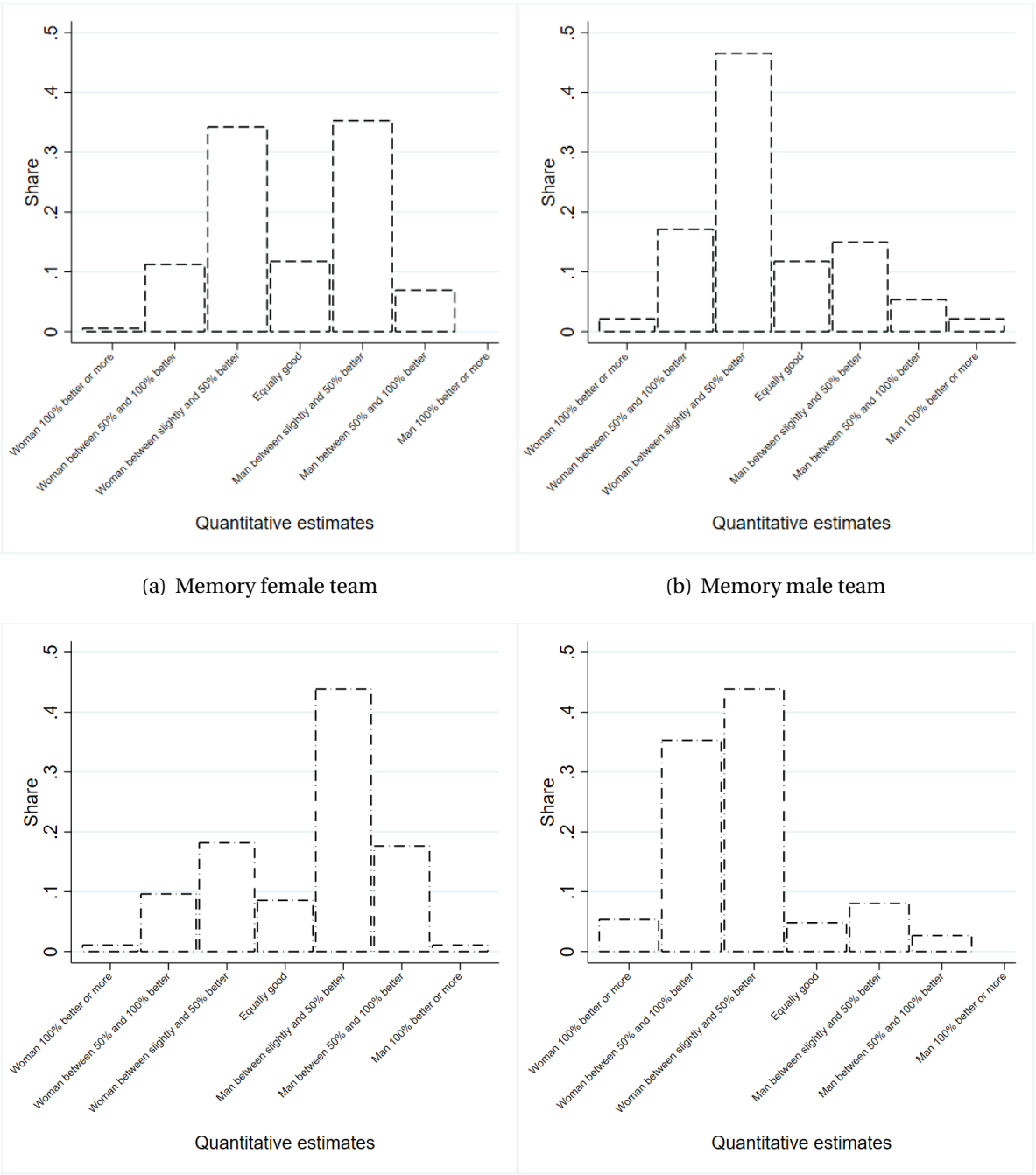
Note: The figure shows the distribution of quantitative assessments for the memory task (dashed black bars) and the shelves task (gray solid bars). For the stereotypically male task (shelves), subjects are more likely to believe that the team with a man will perform better than for the female task (memory).

Figure A.4. Expectations about choices of others



Note: The figure shows the distribution of beliefs about the choices of others for the memory task (dashed bars) and the shelves task (solid bars). For the stereotypical male task (shelves), people are much more likely to believe that others believe the team with a man will perform better.

Figure A.5. Histograms of quantitative estimates



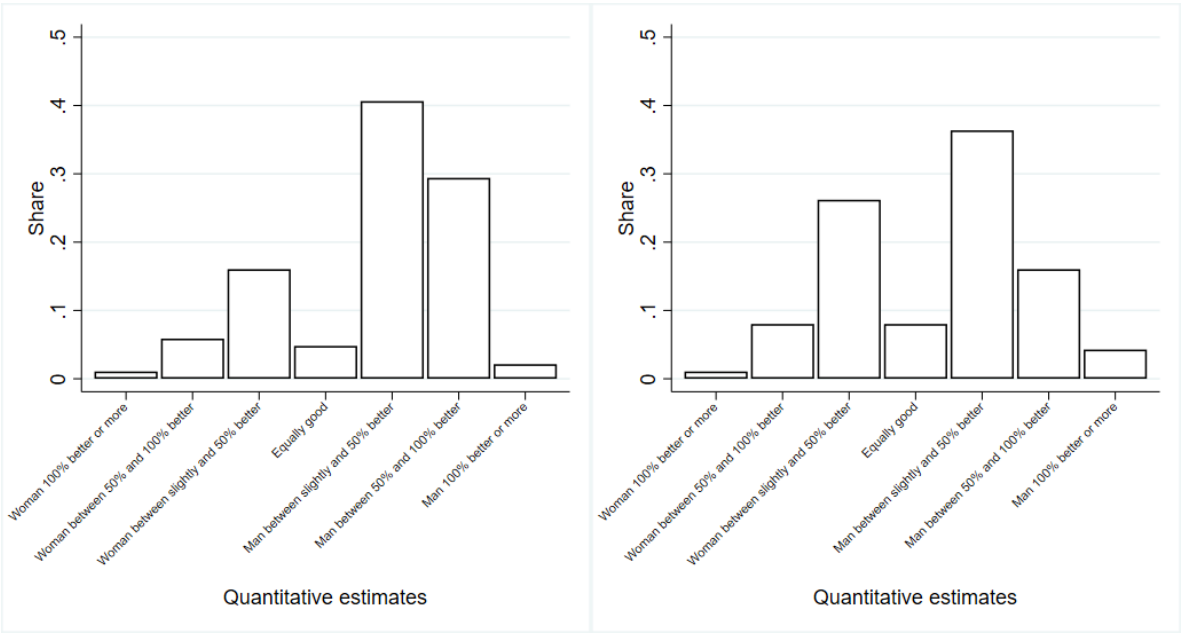
(a) Memory female team

(b) Memory male team

(c) Network female team

(d) Network male team

Figure A.5 (Continued). Histograms of quantitative estimates



(e) Shelves female team

(f) Shelves male team

Note: The figure shows the quantitative assessments by task and initial team composition for the memory task (Panel (a) and Panel (b)), the network task (Panel (c) and Panel (d)), and the shelves task (Panel (e) and Panel (f)). For all three tasks, changing the initial team composition induces a shift in the quantitative assessments: moving from the female teams (panels (a), (c), and (e)) to the male teams (panels (b), (d), and (f)) shifts the distribution to the left, that is, toward a better assessment of the team to which a woman is added.

A.2 Additional tables

Table A.1. Descriptive statistics

	Mean	Sd	p50	Min	Max
Panel A. Student sample ($n = 187$)					
Female	0.508	0.501	1	0	1
Age	22.10	3.996	21	17	50
German	0.882	0.323	1	0	1
Semester	3.743	3.827	3	0	20
Education (in years)	14.73	1.545	14.50	13	18
Field of study					
Natural sciences	0.080	0.272	0	0	1
Maths/Informatics	0.107	0.310	0	0	1
Engineering	0.198	0.399	0	0	1
Business/Economics	0.171	0.378	0	0	1
Humanities	0.193	0.395	0	0	1
Others	0.251	0.435	0	0	1
Panel B. Sample of personnel and general managers ($n = 443$)					
Female	0.503	0.501	1	0	1
Age	46.12	12.07	46	20	82
Education (in years)	13.98	2.698	13	9	17
Job category					
Public official	0.066	0.248	0	0	1
Administration	0.063	0.244	0	0	1
Business man/Accountant	0.086	0.280	0	0	1
Other clerk	0.208	0.406	0	0	1
Manager/Executive	0.160	0.367	0	0	1
No employment	0.061	0.240	0	0	1
Self-employed	0.133	0.340	0	0	1
Other	0.223	0.417	0	0	1
Hiring decisions	2.889	1.294	3	1	5
Attention	5.289	1.665	6	1	7

Note: The table provides mean, standard deviation (Sd), median (p50), minimum value (Min), and maximum value (Max) of individual characteristics. Hiring decisions is a nominal variable between 1 (never) and 5 (very frequently). Attention is a nominal variable between 1 (very inattentive) and 7 (very attentive).

B Robustness analyses

B.1 Regression analysis

In this section, we present a unified analysis of the subjects' bets with respect to the task, the gender composition of the initial team, and demographic variables. We report the results from linear probability models (Table B.1). Probit regressions yield similar results and are reported in Table B.2. Column (1) of Table B.1 shows estimates from a linear probability model of a student choosing a man (instead of a woman) regressed on a dummy for whether the initial team consisted of women, on the task, and on the interactions between an initially all-female team and tasks. We replicate the results from sections 3.2 and 3.3. Task stereotypes are consistent with the choices: the difference in the total share choosing a man between the memory and shelves task is 35 percentage points when the initial team is male (resulting from the difference between 49% and 14%) and 28 percentage points when it is female. Moreover, holding constant the task, the initial gender composition of the team has explanatory power, too, because the difference between an initially all-female and an initially all-male team is significant for all three tasks: 56 percentage points for network, 21 percentage points (56−35 percentage points) for memory, and 15 percentage points (56−41 percentage points) for shelves. The effect on the proportion of men chosen when changing the initial composition is significantly more pronounced for the network task than for the other two tasks. In particular, the probability of choosing a new team member of the opposite gender is significantly more pronounced for the network task that displays complementarities.

Column (2) adds choice-order fixed effects, that is, dummy variables for whether the choice for the given combination of task and initial team composition was made first, second, and so on, and session fixed effects. The coefficients on the five dummies for the second to sixth choice (not reported in the table) are not statistically distinguishable from zero, with one exception.[38] Also, the coefficients of the session dummies are insignificant (not reported). More importantly, the main coefficients of interest are robust to including choice-order and session fixed effects.

In column (3), we add individual-level variables.[39] The coefficients on age, German nationality, semester, and field of study are very small and statistically insignificant. By contrast, being a female student decreases the probability of betting on a man by nine per-

[38] The coefficient for the third choice is positive and significant.

[39] German is a dummy equal to 1 if an evaluator's nationality is German, and 0 otherwise. About 12% of participants are non-German. Semester is a discrete variable between 1 and 20, with mean 3.80. Three subjects are not students. We also control for a subject's field of study by categorizing fields of study into "natural sciences," "maths/informatics," "engineering," "business/economics," "humanities," and "other."

Table B.1. OLS regression of the likelihood of betting on a man

	Students			*Personnel managers*		
	(1)	(2)	(3)	(4)	(5)	(6)
Initial team female	0.561***	0.558***	0.568***	0.300***	0.300***	0.299***
	(0.044)	(0.044)	(0.044)	(0.031)	(0.031)	(0.031)
Memory	0.144***	0.145***	0.148***	0.009	0.009	0.009
	(0.038)	(0.039)	(0.039)	(0.026)	(0.026)	(0.027)
Shelves	0.492***	0.489***	0.487***	0.192***	0.192***	0.193***
	(0.041)	(0.041)	(0.041)	(0.029)	(0.029)	(0.029)
Initial team female×Memory	-0.348***	-0.342***	-0.348***	-0.054	-0.054	-0.054
	(0.063)	(0.064)	(0.065)	(0.038)	(0.038)	(0.039)
Initial team female×Shelves	-0.412***	-0.408***	-0.411***	-0.086*	-0.086*	-0.086*
	(0.060)	(0.060)	(0.061)	(0.041)	(0.041)	(0.042)
Female			-0.095***			-0.063**
			(0.026)			(0.020)
Age			-0.000			-0.001
			(0.005)			(0.001)
German			-0.003			
			(0.045)			
Semester			-0.007			
			(0.005)			
Education						0.003
						(0.004)
Hiring decisions						0.038
						(0.022)
Attention						-0.002
						(0.021)
Constant	0.112***	0.065	0.189	0.226***	0.178***	0.158
	(0.023)	(0.059)	(0.131)	(0.020)	(0.044)	(0.089)
Choice order FE		✓	✓		✓	✓
Session FE		✓	✓		n.a.	n.a.
N	1,122	1,122	1,104	2,658	2,658	2,646

Note: The table shows the results from OLS regressions of betting on the team to which a man is added. Coefficients for field of study are not reported in column (3), and for job category are not reported in column (6). Omitted categories are "initial team male," "network," "task-composition combination evaluated first," "male," "non-German," "natural sciences," and "public official." The lower number of observations in columns (3) and (6) result from three participants not being students and hence not reporting a semester and field of study in the student sample, and two participants indicating an age of five years or "other" education in the manager sample. Heteroskedasticity-robust standard errors clustered at the respondent-level are in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table B.2. Probit regression of the likelihood of betting on a man

	Students			*Personnel managers*		
	(1)	(2)	(3)	(4)	(5)	(6)
Initial team female	0.547***	0.543***	0.557***	0.296***	0.295***	0.295***
	(0.044)	(0.044)	(0.044)	(0.030)	(0.030)	(0.030)
Memory	0.184***	0.185***	0.194***	0.011	0.010	0.011
	(0.049)	(0.049)	(0.049)	(0.031)	(0.031)	(0.031)
Shelves	0.486***	0.483***	0.485***	0.197***	0.196***	0.197***
	(0.042)	(0.041)	(0.042)	(0.030)	(0.030)	(0.030)
Initial team female×Memory	-0.357***	-0.351***	-0.363***	-0.052	-0.052	-0.052
	(0.065)	(0.066)	(0.066)	(0.040)	(0.040)	(0.040)
Initial team female×Shelves	-0.408***	-0.404***	-0.412***	-0.099*	-0.097*	-0.098*
	(0.062)	(0.061)	(0.062)	(0.041)	(0.041)	(0.041)
Female			-0.094***			0.062**
			(0.025)			(0.020)
Age			0.000			-0.001
			(0.005)			(0.001)
German			-0.001			
			(0.045)			
Semester			-0.007			
			(0.005)			
Education						0.003
						(0.004)
Hiring decisions						0.040
						(0.022)
Attention						0.002
						(0.021)
Choice order FE		✓	✓		✓	✓
Session FE		✓	✓		n.a.	n.a.
N	1,122	1,122	1,104	2,658	2,658	2,646

Note: In this table, we reestimate Table B.1. We present marginal effects from probit regressions. The results are similar to the results from the OLS regression. We observe that (i) a stereotypically male task increases the probability of choosing a man (diff. between memory and shelves in the student (manager) sample: 30pp (20pp); $p < 0.001$ each), (ii) the probability of choosing a man is increased if the initial team consists of women, and (iii) this effect is significantly larger for the network task. Effects for field of study are not reported in column (3), and for job category are not reported in column (6). Omitted categories are “initial team male,” “network,” “task-composition combination evaluated first,” “male,” "non-German," “natural sciences,” and “public official.” Heteroskedasticity-robust standard errors clustered at the respondent-level are in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

centage points. Adding the individual-level controls does not affect the main results. In columns (4) to (6), we conduct the same regressions using the sample of general and personnel managers. Column (4) shows that the managers are 18 or 15 percentage points more likely to choose a man for the shelves than for the memory task if the initial team is male or female, respectively. They are also substantially more likely to choose a man if the initial team consists of women, more precisely, 30, 25, and 21 percentage points more likely for the network, memory, and shelves task, respectively. And, the effect of the team composition is more pronounced for the network task than for the other two tasks, although the difference is only marginally significant for the comparison of the network and shelves tasks. In column (5), we add choice-order fixed effects, which leaves the coefficients virtually unchanged. In column (6), we again add individual-level variables.[40] We again find that being a woman decreases the probability of betting on a man, whereas age has no effect on choices. In addition, no partial correlation exists between the managers' educational attainment, their experience in making hiring decisions, and the attention they paid to the experiment and the probability of choosing a man.

Table B.3. Heterogeneity across samples

Interaction	Task		Composition
	Network	Shelves	Init. female
Sample$x$Task	-0.011	-0.149^{***}	
	(0.036)	(0.041)	
Sample$x$Composition			-0.055
			(0.037)

Note: The table presents the results of two OLS regressions of choosing a man on a dummy for whether the initial team consisted of women, on the task, the sample, the interactions between the initially all-female team and the tasks, and (i) the interaction of sample and task (first and second column) or (ii) the interaction of sample and team composition (third column). *,**, and *** denote significance at the 10%, 5%, and 1% level, respectively.

To investigate whether differences exist in the effect of the task stereotype and of the initial team composition between the student and the professional sample, we conduct the regressions of Table B.1 and interact the task or the initial group composition, respectively, with the sample (student or professional). We find that, averaging across the three tasks, the effect of the team composition on subjects' choices is similar for students and

[40] Education is measured in years and varies between 9 and 17 years in the sample. Hiring decisions is a dummy equal to 1 if a manager indicates that they frequently make hiring decisions. Attention is their self-reported attentiveness during the experiment and measured on a 7-point scale.

managers. Averaging across the two team compositions, the effect of changing the task stereotype is significantly more pronounced for the student sample. The effect of changing the task from memory to network is not different for managers and students (see Table B.3).

B.2 Further robustness checks

B.2.1 Classification of evaluator types

We exploit the fact that we observe within-subjects variation in choices and classify evaluators as following task stereotypes or not and preferring homogeneous or diverse teams. We do so based on their choices for the stereotypically female and male task, shelves and memory, for initially all-male and all-female teams. Thus, we use four decisions per evaluator. The results are presented in Figure B.1.[41]

In the student sample, about two-thirds of choices are in line with a preference for diversity, a preference for adding a person consistent with the job stereotype, or both (see the bars with black lines in Panel (a) of Figure B.1).[42] Of the 187 evaluators, 16% follow the stereotype of the job in all four cases, whereas 2% always do the opposite of the task stereotype (and hence do not show a preference for homogeneity or diversity). Thirteen percent always bet on the diverse team (and do not follow the stereotype), compared with only 3% who always bet on the homogeneous team, that is, never on the diverse team. Around 31% show a preference for diversity and follow a stereotype: they either make a diverse choice for the female task (memory) and follow the stereotype for the male task (19%) or make a diverse choice for the male task (shelves) and follow the stereotype for the female task (12%). By contrast, only 12% make choices that are a mix of a preference for homogeneity and task stereotypes. Finally, 4% always choose a man and 5% always choose a woman. Fourteen percent do not fit into any of these classes.[43]

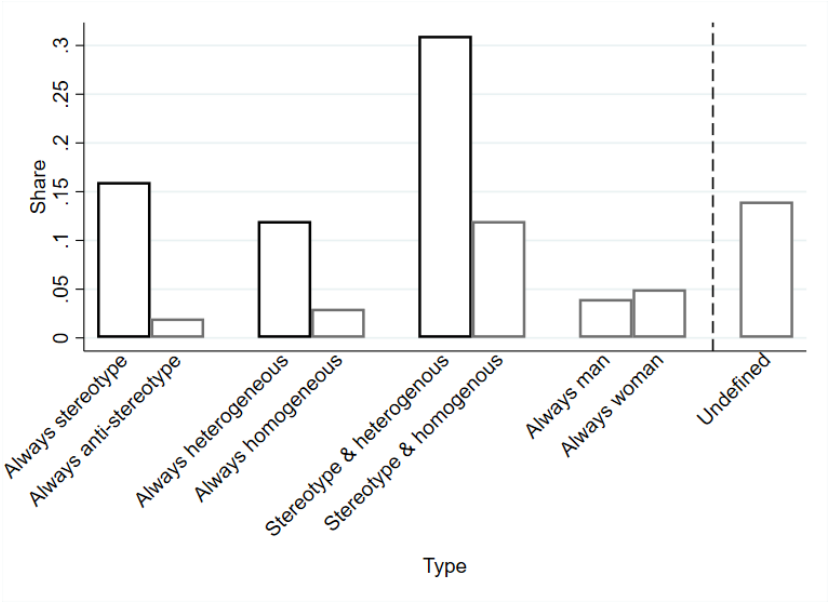
Panel (b) of Figure B.1 reveals a similar picture for managers. In line with some of the results reported above, the data seem to be more noisy in that a larger share of subjects cannot be classified. We presume this noise might be caused by conducting the experiment with the managers online rather than in the lab. Also, they are more likely to always pick a woman. Otherwise, the distribution of types is strikingly similar and the choices of a large share of participants can be explained solely based on a preference for diversity, a preference for adding a person consistent with the job stereotype, or both.

[41] We base this classification on the memory and shelves tasks where we expected clear gender stereotypes. We also analyze how subjects who are classified in this manner choose in the network task. We find, for instance, that the difference in the share of subjects preferring the diverse team for the network task between the initially all-female and the initially all-male team is only 20 or 8 percentage points, respectively, for students and managers who never prefer the diverse team in the memory and shelves task, but is more than three times this amount (65.21 or 60.30 pp) for subjects who always prefer diversity.

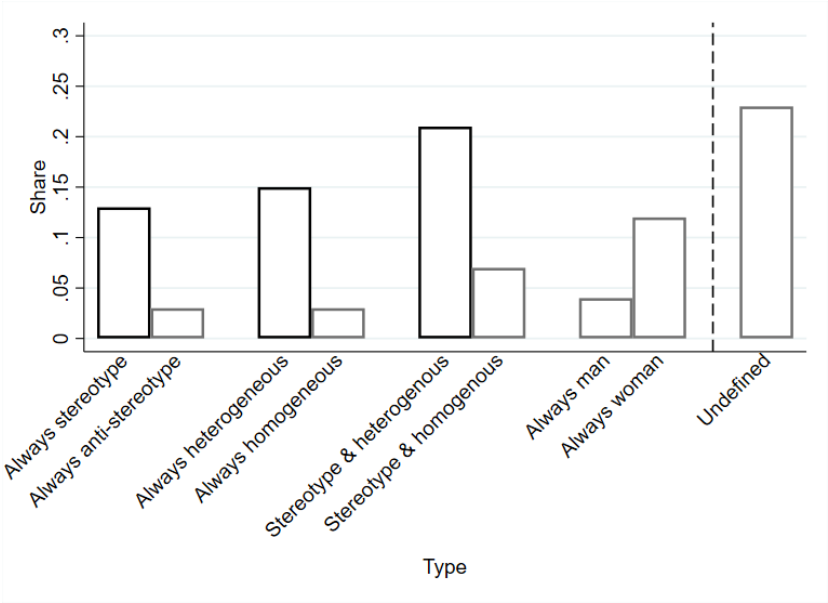
[42] The probability of belonging to these two-thirds does not correlate with any individual characteristics such as age, gender, or field of study.

[43] Because we employ several questions to check the understanding of students, we believe that these decisions are not driven by noise but simply cannot be classified based on (simple) preferences for diversity or homogeneity, for following stereotypes or not, and for betting on a certain gender.

Figure B.1. Classification of evaluators



(a) Students



(b) Managers

Note: The figure shows the classification of evaluator types based on the four decisions for the memory/shelves tasks and the female/male teams for the student sample (Panel (a)) and the manager sample (Panel (b)). The categories are disjoint. The black bars show the share of subject types who make choices consistent with a preference for diversity, a preference for adding a person consistent with the job stereotype, or both.

B.2.2 Robustness analyses laboratory experiment

Table B.4 shows average values of the obtained individual characteristics for the two subject pools of the lab experiment, Berlin and Konstanz, Germany, and p values of tests for differences in means. Gender was successfully balanced for the sessions run in both cities. Some non-negligible differences exist between the subjects from Berlin and Konstanz with respect to other characteristics.[44] In particular, subjects from the Berlin subject pool are older, have studied longer, and are more likely to be non-German. They are also more likely to be studying engineering or related fields, and are less likely to be studying humanities. The latter results reflect the fact that the Berlin subject pool mostly consists of subjects from the Technical University in Berlin, which offers many engineering study programs but almost no humanities.

Table B.4. Descriptive statistics by subject pool

	All	Berlin	Konstanz	p
Female	0.508	0.510	0.505	0.946
Age	22.10	22.78	21.37	0.016
German	0.882	0.792	0.978	0.001
Semester	3.743	4.490	2.956	0.006
Field of study				
Natural sciences	0.080	0.073	0.088	0.706
Maths/informatics	0.107	0.146	0.066	0.077
Engineering	0.198	0.375	0.011	0.001
Business/Economics	0.171	0.125	0.220	0.085
Humanities	0.193	0.063	0.330	0.001
Other	0.251	0.219	0.286	0.291
Observations	187	96	91	

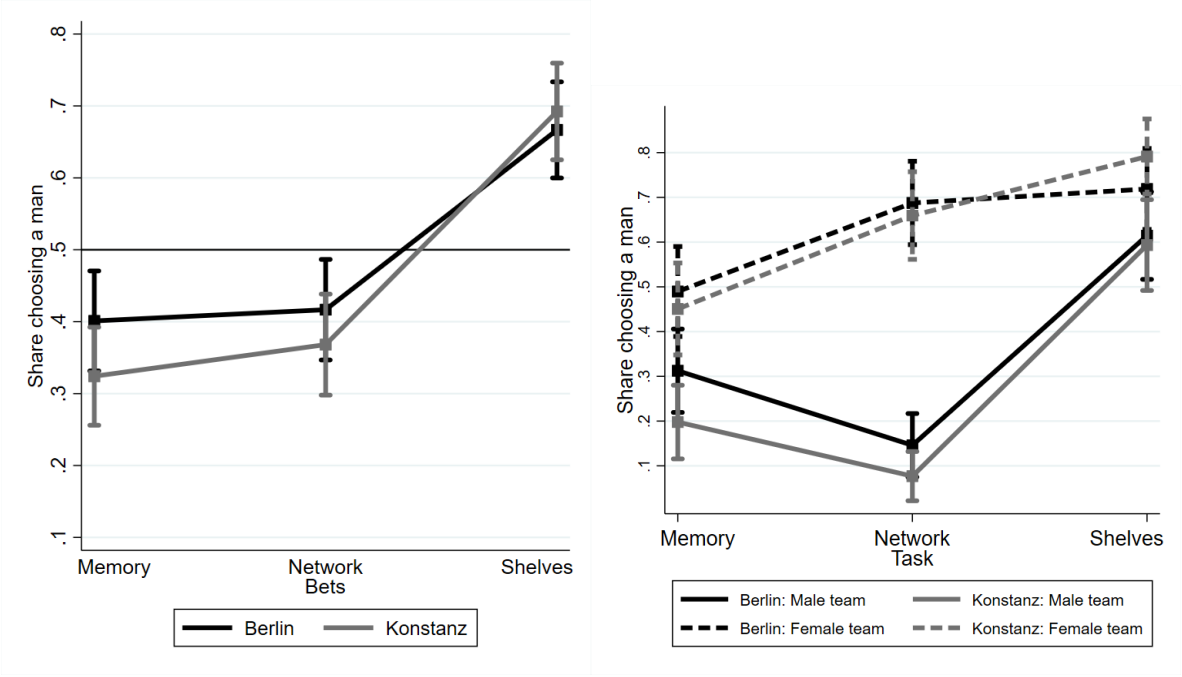
Note: The table shows mean values for individual characteristics by subject pool of the lab experiment, along with p values from z- and t-tests for differences in means between the subject pools for the binary and continuous variables, respectively.

Figures B.2 repeats Figure 2 and 3 of the main text separately by subject pool. The figure reveals that despite the differences in the subject pools, the findings are very similar for both pools. Specifically, Panel (a) of Figure B.2 shows that whereas the difference between the memory and shelves task is slightly smaller in Berlin, in both samples there is a

[44] Konstanz and Berlin are different in various ways. Konstanz is located in Germany's most southwest federal state, Baden-Württemberg, at the boarder to Switzerland. Berlin is a city state located in the northeast of Germany. Konstanz has 83,000 inhabitants, whereas about 3,400,000 people live in Berlin.

substantial difference in choices between the two tasks. Panel (b) of Figure B.2 reveals that subjects in both subject pools believe that diverse teams perform better: in both locations the share choosing a man is higher if the initial team consisted of women. The difference is statistically significant for all tasks in both locations with the exception that it is not statistically significant for the shelves task in Berlin.

Figure B.2. Bets by subject pool



(a) Pooling team compositions

(b) Split by team composition

Note: Panel (a) shows the share of students expecting the team to be more productive when a man is added pooled for both initial composition of teams with 95%-confidence intervals. Panel (b) shows the same separately for each initial team composition.

B.2.3 Robustness analyses online experiment

This appendix presents results from a robustness analysis in which we restrict the professional sample to managers who indicate that they regularly participate in hiring decisions and who self-report that they were attentive when making their choices.[45] The former analysis is informative because it allows us to zoom in on the behavior of our target group, individuals who make hiring decisions in the labor market. The latter is indicated by the experiment being run online. In addition, the heterogeneity of the professional sample in terms of personal characteristics allows us to analyze whether the effects of task stereotype and team composition identified as part of the main analysis vary by age, educational attainment, and occupation.[46]

Figure B.3 presents some graphical evidence. As can be seen in panels (a) to (d) of Figure B.3, the effects of task stereotype and team composition are similar for managers (a) who were attentive during the experiment or inattentive, (b) who regularly make hiring decisions or not, (c) who have above or below median age, and (d) who have above or below median educational attainment.[47] This holds particularly true for the effect of the task stereotype. Small differences exist in the degree to which the subjects believe in the superiority of diverse teams. For instance, Panel (b) shows that for less attentive subjects (gray lines), the preference for diversity in the network task is less pronounced, indicated by a smaller difference between the dashed and solid lines, than for attentive subjects (black lines). This finding seems plausible. In addition, we perform the regressions described in section 3.5 and formally analyze the interaction of these variables with the task and the initial group composition (see Table B.5). The analyses support the graphical evidence. In particular, we find that neither a subject's experience in hiring decisions, attentiveness during the experiment, nor educational attainment significantly moderates the effect of task (network vs. memory and shelves vs. memory, respectively) and team composition. We find some evidence that for individuals working as a business person/accountant, the effect of the shelves task is smaller ($p = 0.001$, not reported in the Table). Overall, the effects of the task and composition are therefore robustly observed across different subsamples.

[45] In our preanalysis plan, we specified to also restrict the analysis to those who fully answered the questionnaire questions. However, as described in the addendum to our plan (see section D), we conducted the experiment in collaboration with a data-collecting agency and then decided to require all participants to complete the questionnaire.

[46] We analyze how gender moderates the observed effects as part of the main analysis (see section 3.4).

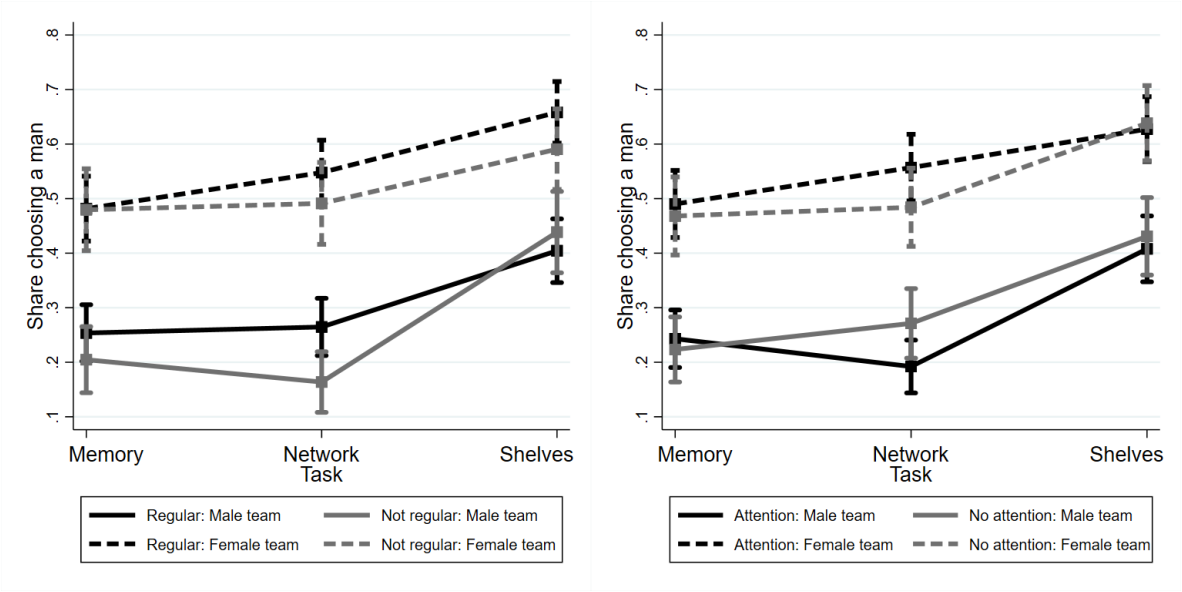
[47] We preregistered to split the sample at the median with respect to how regularly people make hiring decisions and analyze the half that makes hiring decisions more regularly as well as to analyze the effects for those who have participated at least "occasionally." In our sample, both categorizations lead to the same individuals being classified as making hiring decisions regularly. Equally, we preregistered to split the sample at the median with respect to how attentive participants were during the experiment. For the age and education variables, the median splits are exploratory. For the regressions below, we use the continuous educational attainment variable.

Table B.5. Heterogeneity across samples

Interaction	Task		Composition
	Network	Shelves	Init. female
Hiring decisions			
Hiring decisions$x$Task	−0.053	−0.009	
	(0.041)	(0.047)	
Hiring decisions$x$Composition			0.003
			(0.041)
Attentiveness			
Attentiveness$x$Task	−0.024	−0.038	
	(0.042)	(0.047)	
Attentiveness$x$Composition			0.056
			(0.041)
Age			
Age$x$Task	0.017	0.019	
	(0.041)	(0.047)	
Age$x$Composition			0.045
			(0.040)
Education			
Education$x$Task	−0.003	−0.002	
	(0.007)	(0.008)	
Education$x$Composition			0.010
			(0.007)

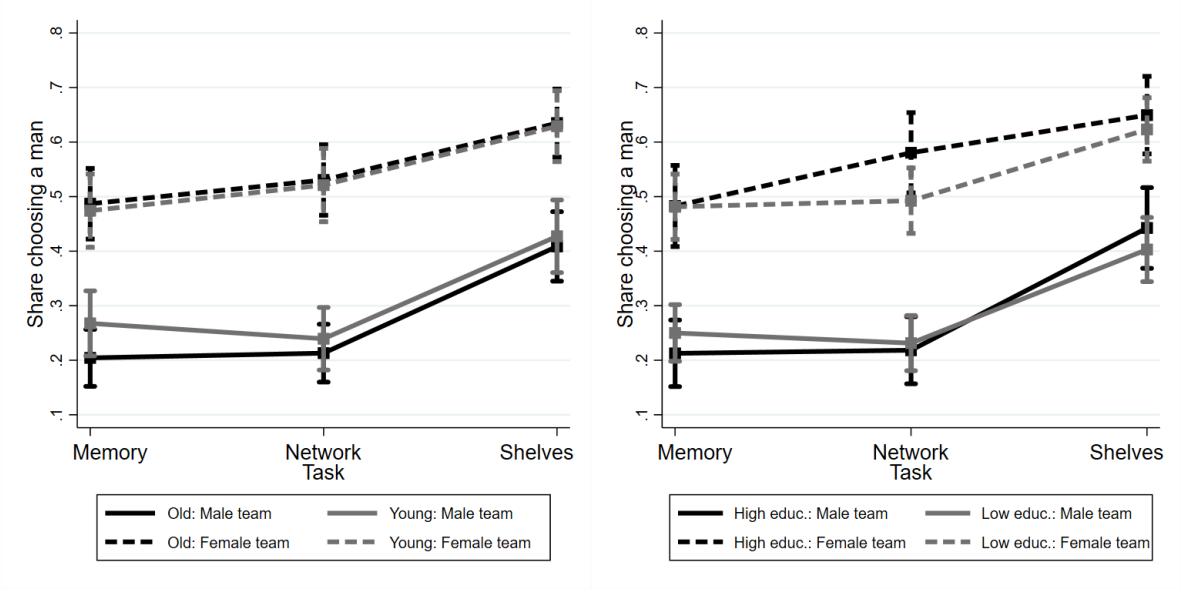
Note: The table presents the results of eight OLS regressions of choosing a man on a dummy for whether the initial team consisted of women, on the task, the respective moderator variable of interest, the interactions between an initially all-female team and tasks, and (i) the interaction of task and the moderator (first and second column) or (ii) the interaction of team composition and the moderator (third column). As preregistered, we use dummy variables for managers who have above-median experience with hiring decisions or report to have been above-median attentive. For age, we use a dummy variable for above-median age. For educational attainment, we recode the discrete education variable and the new variable takes on values between 9 and 18 years. Not reported are the coefficients of the interactions with subjects' job title. *,**, and *** denote significance at the 10%, 5%, and 1% level, respectively.

Figure B.3. Heterogeneity analysis



(a) Hiring decisions

(b) Attentiveness



(c) Age

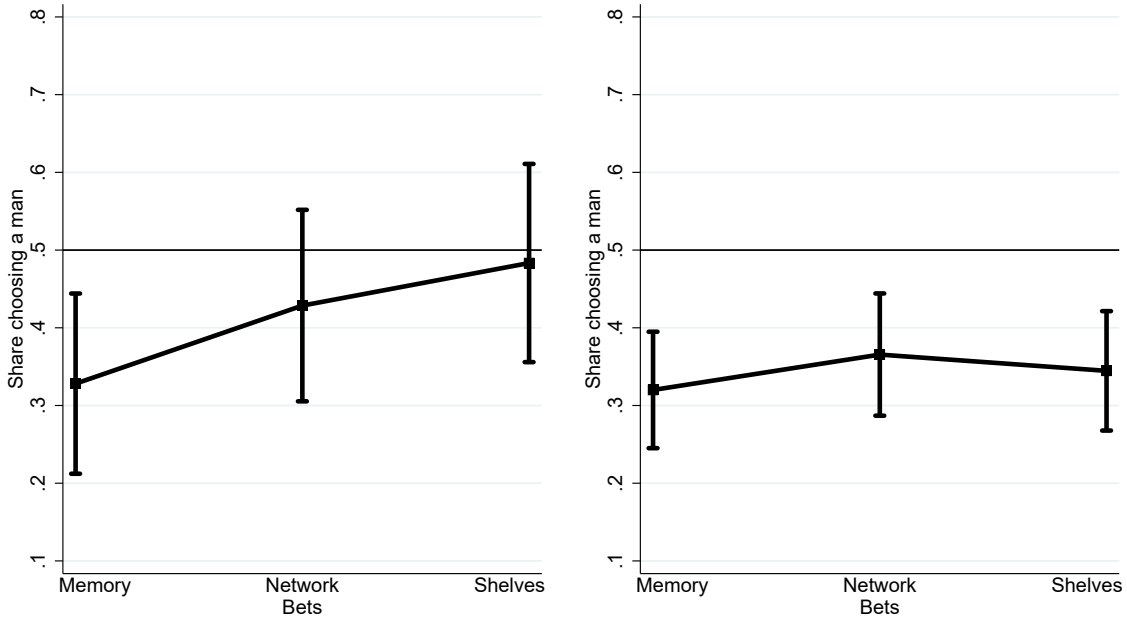
(d) Education

Note: The figure shows the share betting on a man for managers who (i) regularly make hiring decisions (black lines in Panel (a)) or not (gray lines in Panel (a)), (ii) were attentive during the experiment (black lines in Panel (b)) or not (gray lines in Panel (b)), (iii) are of above median age (black lines in Panel (c)) or below (gray lines in Panel (c)), and (iv) obtained more than the median education (gray lines in Panel (d)) or less (gray lines in Panel (d)), along with 95%-confidence intervals. In all panels, solid lines refer to initially all-male teams and dashed lines to initially all-female teams.

B.2.4 Order effects

In Figure B.4 and Figure B.5 below, we present our main findings, looking at the first choices that participants make only. Note that whereas the regression results reported in section 3.5 control for the order in which subjects make their choices, here we restrict the sample to subjects' first choice. Figure B.4 shows that for both the lab experiment with students (Panel (a)) and the online experiment (Panel (b)) the effect of the task stereotype is attenuated when we restrict the sample to the first choice that the subjects make. Specifically, the share of subjects choosing a man in the shelves task is lower than when analyzing all of subjects' choices. This finding is particular pronounced in the manager sample, where we find no difference in the share choosing a man between the memory and shelves task, looking at first choices only.

Figure B.4. Bets by task between-subjects variation



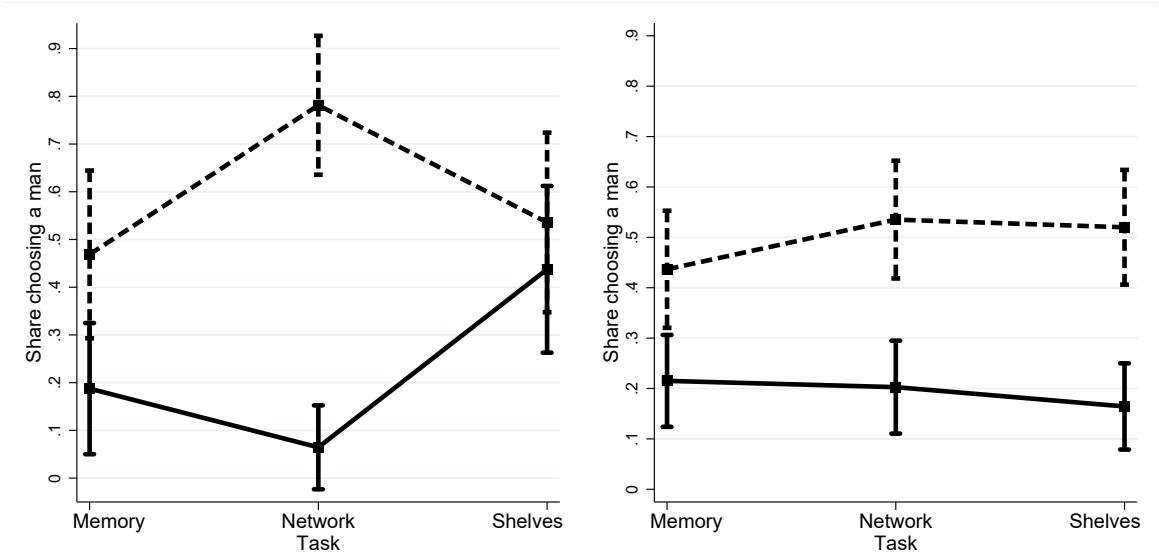
(a) Students (b) Managers

Note: The figure shows the share of students (Panel (a)) and managers (Panel (b)) betting on a man together with 95%-confidence intervals when we analyze the first choice that participants make.

We can only speculate about the reasons for this finding. Carlsson et al. (2012) provide evidence of increased probabilities of making errors in early choices. This explanation would be in line with the finding that our results are more robust for the student sample. Remember that in the lab experiment, we included control questions to check for subjects' understanding, whereas we did not include these questions in the online exper-

iment. Alternatively, subjects might behave differently in later choices, due to cues about what constitutes appropriate behavior in the experiment (Zizzo, 2010). In this sense, we may underestimate the effect of the task stereotype on subjects' beliefs analyzing only first choices, because the bets may be influenced by descriptive norms about the socially desirable behavior of not following the male stereotype. A third potential explanation for the finding is based on anecdotal evidence suggesting that whether women or men are better at assembling shelves is unclear; therefore, the stereotype for the shelves task may be less explicit. For instance, in 2008, the CEO of IKEA Germany stated that women are more skilled than men at assembling the company's furniture (see The Sydney Morning Herald, 2008; Wiking et al., 2016, provide evidence refuting this claim). The subjects therefore may only form the stereotype for the shelves task in the face of the other tasks. Note, however, that based on our stereotype elicitation conducted at the end of the online experiment, we find a clear stereotype for the shelves task (see section 3.1). Also note that in the lab experiment, the differences in choices between the tasks are still statistically significant even if we look at first choices.

Figure B.5. Bets by task and group composition between-subjects variation



(a) Students (b) Managers

Note: The figure shows the share of students (Panel (a)) and managers (Panel (b)) betting on a man with 95%-confidence intervals when we analyze the first choice that participants make. The solid lines refer to initially all-male teams and the dashed lines to initially all-female teams.

In both samples, we observe the belief that diverse teams are more productive even if we only look at first choices (see Figure B.5). The difference in the share choosing a man

is highly significant if we pool choices across tasks. It is also highly significant for all three tasks in the online experiment and two out of three tasks in the lab experiment, considering the tasks separately.

C Experimental instructions

C.1 Laboratory experiment with students

The original instructions were in German. We document the instructions of evaluators that include the descriptions of the tasks in the team sessions. The parts in square brackets were not shown to the participants. We randomized whether evaluators make their first three predictions for an initially all-female or all-male team and the order of tasks within these two blocks. Below, we show the instructions of an evaluator who first makes predictions for an initially all-male team for the shelves, memory, and network task, respectively, and then for an initially all-female team.

Screen 1:
Welcome to our experiment!

The purpose of today's experiment is to study economic decision-making. For your participation today, you will receive a fixed amount of €7. You can earn additional money depending on your decisions.

During the experiment, you are not allowed to use any electronic devices. Please only use the programs that are part of the experiment. Please do not talk to other participants. If you have a question, please raise your hand. We will then come to you and answer your question in person. Do not pose the question out loud. If the question is relevant for all participants, we will repeat it and answer it for everyone. Should you violate any of these rules, we will have to exclude you from the experiment and the payment.

In this experiment, we ask you to predict how well teams that consist of four members perform different tasks. Your payment depends on the accuracy of your predictions.

Please note: The teams will solve the tasks only in the following days. Because your payment depends on the performance of those teams, you will receive your payment only in three weeks. You can then either collect your payment personally or receive an Amazon coupon via email, or we can transfer the payment directly to your bank account. You can tell us at the end of the experiment how you want to receive your payment, and provide us with the necessary information. At the end of the experiment, you will also receive the fixed amount in cash.

Screen 2:
In the following, you will read a description of the task performed by two teams consisting of **four male participants**. Here are the instructions that those teams receive:

"Description of the task in Part 1

Your task is to assemble two IKEA LAIVA bookshelves. You are allowed to use the included instructions and work on both bookshelves at the same time. The bookshelves have to be fully built such that they are steady and ready for use. That means that individual pieces may not be loose, and that the shelves have to be stable enough to hold the weight of books or stable enough to be carried. We will verify whether the bookcases are steady and ready to use as soon as you are done.

Your goal is to assemble the bookcases as fast as possible. Your payment at the end of the experiment will be calculated as follows: You will receive €100 per person divided by the minutes that you needed to complete the task. For example, if your team needs 10 minutes to assemble the bookshelves, then each member will receive €10. If you do not manage to assemble the bookshelves in 40 minutes, the task will be terminated and each member will receive €2.50."

After the participants have completed this task, one randomly selected member of the team will be replaced by a new person. In one of the teams, the new person will be a man, and in the other team, the new person will be a woman. Which of the two teams a man or woman is allocated to is randomly determined. To determine who leaves the team, the experimenter rolls a die. If he rolls a 5 or a 6, he will have to roll again. If he rolls a number between 1 and 4, the person whose ID corresponds to that number will have to leave the team. This randomly selected person receives her payment, and the experiment is over for them.

Screen 3:
For the second part of the experiment, the new participant enters the laboratory. This participant, like the others, has received the instructions for the first part of the experiment, along with the description of the task. The participant has been waiting in the adjoining room and receives the same payment as the other members of the team for part 1 of the experiment.

Afterwards, the team consisting of the three initial members and the new member perform a new task. As before, the performance of the team in this task determines the amount of money the participants receive in the experiment.

In the following, you can read the instructions the participants receive for the second part of the experiment:

"Description of the task in Part 2

Your task is now to once again assemble two bookshelves, this time, however, IKEA BILLY bookshelves. All rules from the first part of the experiment remain valid.

Once more, it is your goal to assemble the bookshelves as quickly as possible. Your payment

at the end of the experiment is calculated as before: You will receive €100 per person divided by the minutes that you needed to complete the task. For example, if your team needs 10 minutes to assemble the bookshelves, then each member will receive €10. If you do not manage to assemble the bookshelves in 40 minutes, the task will be terminated and each member will receive €2.50."

> **Your task:** We now ask you to predict the performance of the two teams in the second part of the experiment. Reminder: Both teams consist of four men in part 1 of the experiment. Afterwards, the team in which a man will be replaced by another man will be randomly determined, as will the team in which a man will be replaced by a woman. You are now asked to tell us whether you believe that the team with four men will do better or worse than the team with three men and one woman. If your prediction is correct, you will receive **€1.20**. Both options are considered correct if the teams perform equally.

Screen 4:

Before you make your predictions, we ask you to answer a few questions to check your understanding. The purpose of these questions is to eliminate possible misunderstandings. You will only be able to advance to the next screen after all questions have been answered correctly. By pressing "back" you can go back to the previous screens. If you have any questions, please raise your hand. We will then come to you and answer your questions in person.

1. How many people are in each team in part 1 of the experiment?
2. How many people are in each team in part 2 of the experiment?
3. How many of those who participated in part 1 also perform the task in part 2 (per team)?
4. How many men perform the task in part 1 per team?
5. How many men perform the task in part 2?
   Please specify both possibilities.
6. How many women perform the task in part 1 per team?
7. How many women perform the task in part 2?
   Please specify both possibilities.

8. For which part do we ask you to compare and predict the performances of the two teams? (Please enter either "1" or "2," depending on the part you think is right.)

9. Please answer another question regarding the task of the teams. Please tell us which of the following statements is right:
   a. The team that assembles the bookshelves faster performs better than the other team.
   b. The team that assembles the prettier bookshelves performs better than the other team.

Screen 5:

Please make your prediction now. If your prediction is correct, you will receive €1.20.

- The team with four men will perform better in part 2 than the team with three men and one woman in part 2.
- The team with three men and one woman will perform better in part 2 than the team with four men in part 2.

Screen 6:[48]

Please now also tell us how big you expect the difference in performance of the two teams in part 2 to be.

- The team with four men will perform at least two times better in part 2 than the team with three men and one woman.
- The team with four men will perform between two times and 50% better in part 2 than the team with three men and one woman.
- The team with four men will perform between 50% faster and slightly better in part 2 than the team with three men and one woman.
- The team with four men will perform as well as the team with three men and one woman (within one minute).

[48] Here and below, we show you the choice options for someone who on the previous screen predicted that the team with four men would perform better. In case someone predicted three men and one woman would perform better, the choice options are analogous.

Screen 7:

We now again ask you to make a prediction, but this time for another task that two new teams have to solve. In the following, you will read a description of the task performed by two teams consisting of **four male participants**. Here are the instructions that those teams receive:

"Description of the task in Part 1

In this part of the experiment, you are asked to solve a memory game. The goal of the game is to find all pairs of cards with the same picture. You can flip two cards in a row by clicking on them on the screen. By clicking those cards again, they turn back over again in case they do not show the same picture. If you are able to flip two cards with the same picture in a turn, this pair will be automatically removed from the game. You can discuss within your team which cards you want to flip next.

Your task is to solve the memory game with your team with the fewest number of turns possible. A turn consists of flipping two cards that do not show the same picture. Your payment will amount to €100 per person if you are able to solve the game with five turns or less. If you need more than five turns to solve the memory, you will receive €500 per person, divided by the number of turns that you needed to complete the game. For example, if your team needs 50 turns to solve the memory game, each member will receive €10. If you do not manage to solve the game with 200 turns or less, the task will be terminated and each member will receive €2.50."

After the participants have completed this task, one randomly selected member of the team will be replaced by a new person. In one of the teams, the new person will be a man, and in the other team, the new person will be a woman. To which of the two teams a man or woman is allocated, is randomly determined. To determine who leaves the team, the experimenter rolls a die. If he rolls a 5 or a 6, he will have to roll again. If he rolls a number between 1 and 4, the person whose ID corresponds to that number will have to leave the team. This randomly selected person receives their payment, and the experiment is over for them.

Screen 8:

For the second part of the experiment, the new participant enters the laboratory. This participant, like the others, has received the instructions for the first part of the experiment, along with the description of the task. The participant has been waiting in the adjoining room and receives the same payment as the other members of the team for part 1 of the experiment.

Afterwards, the team consisting of the three initial members and the new member perform a new task. As before, the performance of the team in this task determines the amount of

money the participants receive in the experiment.

In the following, you can read the instructions the participants receive for the second part of the experiment:

"Description of the task in Part 2

Your task is now to once again solve the memory game with the fewest number of turns possible. A turn consists of flipping two cards that do not show the same picture. Your payment at the end of the experiment is calculated as before: It will amount to €100 per person if you are able to solve the game with five turns or less. If you need more than five turns to solve the memory, you will receive €500 per person, divided by the number of turns that you needed to complete the game. For example, if your team needs 50 turns to solve the memory game, each member will receive €10. If you do not manage to solve the game with 200 turns or less, the task will be terminated and each member will receive €2.50."

Your task: We now ask you to predict the performance of the two teams in the second part of the experiment. Reminder: Both teams consist of four men in part 1 of the experiment. Afterwards, the team in which a man will be replaced by another man will be randomly determined, as will the team in which a man will be replaced by a woman. You are now asked to tell us whether you believe that the team with four men will do better or worse than the team with three men and one woman. If your prediction is correct, you will receive **€1.20**. Both options are considered correct if the teams perform equally.

Screen 9:

We now ask you to answer a question regarding the task of the teams to check your understanding. You will only be able to advance to the next screen after the question has been correctly answered. By pressing "back" you can go back to the previous screens.

Please tell us which of the following statements is true:

a. The team that solved the memory game faster performs better than the other team.

b. The team that solved the memory game in fewer turns performs better than the other team.

Screen 10:

Please make your prediction now. If your prediction is correct, you will receive €1.20.

- The team with four men will perform better in part 2 than the team with three men and one woman in part 2.
- The team with three men and one woman will perform better in part 2 than the team with four men in part 2.

Screen 11:

Please now also tell us how big you expect the difference in performance of the two teams in part 2 to be.

- The team with four men will perform at least two times better in part 2 than the team with three men and one woman.
- The team with four men will perform between two times and 50% better in part 2 than the team with three men and one woman.
- The team with four men will perform between 50% faster and slightly better in part 2 than the team with three men and one woman.
- The team with four men will perform as well as the team with three men and one woman.

Screen 12:

We now again ask you to make a prediction, this time about another task that again has to be performed by two new teams. In the following, you will read a description of the task performed by two teams consisting of **four male participants**. Here are the instructions that those teams receive:

"Description of the task in Part 1

Your task is to draw a network on the paper that we provided. The network should consist of TV series, TV shows, and movies on one side, and actors/actresses on the other. This means that you should write down the name of an actor/actress and then connect the name with a line to a series, show, or movie in which the actor/actress participated. Then, you connect the series, show, or movie with another actor/actress that participated in it, etc. Each actor/actress can be connected to a series, a show, or a movie, but not to another actor/actress.

Equally, each series, show, or movie can be connected to an actor/actress, but not to another series, show, or movie. However, a series, show, or movie can be connected to multiple actors/actresses, and an actor/actress can be connected to multiple series, shows, or movies. Each series, show, or movie and each actor/actress can only be used once in the network.

You as a team are allowed to verify a connection once per minute. This means you are allowed to ask the experimenter whether an actor/actress actually participated in a series, show, or movie or not. The experimenter will verify the connection as quickly as possible, and inform the team whether the connection is correct or not. If the connection is false, you are allowed to erase it and replace it with a new one.

Your goal is to build the longest possible chain. To determine the longest chain, we will count the number of connections after the time has run out. Should a connection be false, that is, if an actor/actress did not participate in the series, show, or movie to which they are connected, the chain is "interrupted" at that particular connection, and the longest remaining chain will be considered. Spelling errors are not a problem, as long as the actor/actress or the series, show, or movie can still be identified. For actors/actresses, mentioning their last name is sufficient. For movies, you have to correctly name the German title or the original title. A movie is considered not yet used when the title of the movie changes by more than one number. For example, if you used the movie "Fack ju Göthe," you cannot use the movie "Fack ju Göthe 2" anymore. If we cannot identify an actor/actress, or if the title of the series, show, or movie is not correct, the network is "interrupted" at this particular point.

You have 10 minutes to solve the task. Your payment at the end of the experiment will be calculated as follows: If the team has made Y correct connections after the 10 minutes have gone by, each team member receives €0.4*Y. You therefore receive 40 cents for each correct connection in the longest chain. For example, if the team created 25 correct connections, each member receives €10. Please note that only the longest chain is relevant for your payment, and not the total number of connections.

After the participants are done with this task, one randomly selected member of the team will be replaced by a new person. In one of the teams, the new person will be a man, and in the other team, the new person will be a woman. To which of the two teams a man or woman is allocated, is randomly determined. To determine who leaves the team, the experimenter rolls a die. If he rolls a 5 or a 6, he will have to roll again. If he rolls a number between 1 and 4, the person whose ID corresponds to that number will have to leave the team. This randomly selected person receives her payment, and the experiment is over for them.

Screen 13:

For the second part of the experiment, the new participant enters the laboratory. This par-

ticipant, like the others, has received the instructions for the first part of the experiment, along with the description of the task. The participant has been waiting in the adjoining room and receives the same payment as the other members of the team for part 1 of the experiment.

Afterwards, the team consisting of the three initial members and the new member perform a new task. As before, the performance of the team in this task determines the amount of money the participants receive in the experiment.

In the following, you will read the instructions the participants receive for the second part of the experiment:

"Description of the task in Part 2

Your task is now to once again build a network. All the rules from the first part remain valid. However, you are not allowed to use connections that were part of the longest chain in the first part anymore. That means that if you connected a certain actors/actress with a certain series, show, or movie, you cannot use this connection again. If you use one of those connections again, the chain will be considered interrupted at that particular point. You are, however, allowed to connect those same actors/actresses with other series, shows, or movies, or the same series, show, or movies with another actors/actress. You may also again use connections that were not part of the longest chain in the first part. You are allowed to still see the network you produced during the first part.

*Your goal is therefore to again build the longest possible chain, and you again have 10 minutes to achieve this goal. Your payment will be calculated as in the first part: If the team has made Y correct connections after the 10 minutes have gone by, each team member receives €0.4*Y. You therefore receive 40 cents for each correct connection in the longest chain. For example, if the team created 25 correct connections, each member receives €10.*

Your task: We now ask you to predict the performance of the two teams in the second part of the experiment. Reminder: Both teams consist of four men in part 1 of the experiment. Afterwards, the team in which a man will be replaced by another man will be randomly determined, as will the team in which a man will be replaced by a woman. You are now asked to tell us whether you believe that the team with four men will do better or worse than the team with three men and one woman. If your prediction is correct, you will receive **€1.20**. Both options are considered correct if the teams perform equally.

Screen 14:

We now again ask you to answer a question regarding the task of the teams to check your

understanding. You will only be able to advance to the next screen after the question has been correctly answered. By pressing “back” you can go back to the previous screens.

Please tell us which of the following statements is true:

a. The team that wrote down the most actors/actresses performs better than the other team.

b. The team that wrote down the longest chain of connections between actors/actresses and series, shows, or movies performs than the other team.

c. The team that wrote down the most series, shows, or movies performs better than the other team.

Screen 15:

Please make your prediction now. If your prediction is correct, you will receive €1.20.

- The team with four men will perform better in part 2 than the team with three men and one woman in part 2.
- The team with three men and one woman will perform better in part 2 than the team with four men in part 2.

Screen 16:

Please now also tell us how big you expect the difference in performance of the two teams in part 2 to be.

- The team with four men will perform at least two times better in part 2 than the team with three men and one woman.
- The team with four men will perform between two times and 50% better in part 2 than the team with three men and one woman.
- The team with four men will perform between 50% faster and slightly better in part 2 than the team with three men and one woman.
- The team with four men will perform as well as the team with three men and one woman.

Screen 17:

In the following screens, you will be again asked to make predictions regarding teams that work on the three tasks; this time, however, the composition of the teams will differ.

Screen 18-29:
[Participants make predictions for the three tasks, this time with the other initial team composition (i.e., with the team initially consisting of four women if the first three decisions were for teams consisting initially of four men – as depicted here – and vice versa).]

Screen 30:
Finally, we would like to ask you a few additional questions: Please try to estimate the predictions that the other participants in this room made concerning the different situations. More precisely, we ask you to estimate, for each of the six constellations you previously decided on, how many out of 11 randomly selected participants in this room predicted that a certain team performs better in part 2 than the other.

The payment you receive for answering these questions will be calculated in the following way: For each question that you answer correctly, you will receive €1.20.

Screen 31:
Task: Assembling bookshelves

Part 1: four men; Part 2: four men versus three men and one woman

Your estimation concerning the prediction of 11 randomly selected other participants in the experiment:

- Number of predictions that in part 2 the team **with four men** will perform better than the team with **three men and a woman**
- Number of predictions that in part 2 the team with **three men and one woman** will perform better than the team **with four men**

Screen 32-36:
[Participants make predictions for the other five constellations in the same way as on Screen 31.]

Screen 37:
Thank you for your predictions! We now ask you to share some personal information with us. Afterwards, you will be able to tell us how you want to receive your payment and you will receive the fixed payment of €7 in cash.

Screen 38:
Please share with us the following personal information:

Please choose your gender.

How old are you?

What is your nationality?

What is your field of study? If you do not study, please tell us what your current occupation is.

In which university semester are you? If you are not a student, please enter “0.”

C.2 Online experiment with managers

Parts in square brackets were not shown to the participants. Original instructions were in German. We randomized whether evaluators make their first three predictions for an initially all-female or all-male team. We once randomized the order of tasks but then kept the order for all blocks of the experiment.

Screen 1: Which teams are more productive?
In this survey, you are asked to assess how well teams perform different tasks. We actually had the teams work on the different tasks, and we measured how well they performed. We now ask you to guess which teams did better.

The better you guess, the higher your potential payout: In this way, you can win more than 1200 Mingel points, which you will be credited via the platform.

By clicking on "Continue," you agree to participate in this survey. You can find details about the consent form *here*.

Screen 2: Assembling shelves
We asked two teams, initially consisting of four male [female] participants, to each assemble two LAIVA shelves from IKEA as quickly as possible. Then, we replaced in one team one member with a man [woman] and in the other team one member with a woman [man].

Both teams then had the task of assembling two shelves of the type BILLY from IKEA in the 2nd part. Once again, the teams received a higher payout the faster they completely built the two shelves. You can find further details *here*.

Which of the two teams needed less time to assemble the two shelves in part 2?

Click on the team. If your guess is correct, you will receive an additional 120 Mingel points.

[Images to make choice.]

Screen 3: Memory game
We also asked two other teams, initially consisting of four male [female] participants, to each solve a memory game with as few mistakes as possible. Then, we replaced in one team one member with a man [woman] and in the other team one member with a woman [man].

Both teams then had the task of solving a new memory in the 2nd part. Again, the teams received a higher payout the fewer mistakes they made. You can find further details *here*.

Which of the two teams made fewer mistakes before solving the memory in part 2?

Click on the team. If your guess is correct, you will receive an additional 120 Mingel points.

[Images to make choice.]

Screen 4: Network of movies and actors
Finally, we asked two other teams, initially consisting of four male [female] participants, to each draw the longest possible network that alternates between TV series, shows, and movies on the one hand, and actors/actresses on the other. Then, we replaced in one team one member with a man [woman] and in the other team one member with a woman [man].

Both teams then had the task of drawing a new network in the 2nd part. Once again, the teams received a higher payout the longer the chain of actors and series, show, or movie they drew. You can find further details *here*.

Which of the two teams drew a longer network in part 2?

Click on the team. If your guess is correct, you will receive an additional 120 Mingel points.

[Images to make choice.]

Screen 5: Further assessments
We now ask you to make some further assessments: Again, we had teams work on the three tasks and measured how well they performed. Unlike before, you now choose between two teams, both of which initially consisted of four female [male] participants, and of which we replaced one member with a man [woman] in one team and one member with a woman [man] in the other.

Assembling shelves:

The teams had the task to assemble two shelves as quickly as possible (*details*).

...

Screen 6: Further assessments
Memory game:

The teams had the task to solve a memory game with as few mistakes as possible (*details*).

...

Screen 7: Further assessments
Network of movies and actors:

The teams had the task to draw a network as long as possible (*details*).

...

Screen 8: What do others believe?
A few more questions: How do you think other participants assessed the situations?

Building shelves (*details*):

Out of 11 other participants, how many believe that the team with **four men [women]** was better than the team with **three men [women] and one woman [man]**?

We will compare your answer with the actual assessments of 11 other participants. For the correct answer, you will receive an additional 60 Mingel points.

[Buttons for 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 to make choice].

Out of 11 other participants, how many believe that the team with **four women [men]** was better than the team with **three women [men] and one man [woman]**?

We will compare your answer with the actual assessments of 11 other participants. For the correct answer, you will receive an additional 60 Mingel points.

[Buttons for 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 to make choice].

Screen 9: What do others believe?

...

Screen 10: What do others believe?

...

Screen 11: And who was better in part 1?

Last but not least: On average, was the team consisting of four men or the team consisting of four women better in part 1 in the different tasks?

Assembling shelves:

Which of the two teams needed less time to assemble the two shelves in part 1?

Click on the team. If your guess is correct, you will receive an additional 60 Mingel points.

[Images to make choice.]

Screen 12: And who was better in part 1?

...

Screen 13: And who was better in part 1?

...

Screen 14: A few short questions at the end ...

Please choose your gender (female, male, diverse):

How old are you?

What is your current occupation?

What is your highest level of education? (no degree, lower secondary degree, intermediary secondary degree, highest secondary degree/Abitur, vocational training, technical col-

lege/university, other)

How often have you participated in hiring decisions or recruiting processes in the past? (never, rarely, occasionally, frequently, very frequently).

Screen 15: A few short questions at the end ...
Would you say you answered the questions and assessments in this survey rather inattentively or attentively? Please answer the question as honestly as possible, there is no disadvantage for you if you answered the questions less attentively. (very inattentive, inattentive, rather inattentive, neutral, rather attentive, attentive, very inattentive)

Screen 16: End of survey
Thank you for your participation in this survey!

Based on your assessments, you will receive an additional payout of X Mingel points.

You have the opportunity for anonymous feedback here:

Alternatively, you can contact us at d.kuebler@wzb.eu for questions, feedback, or criticism.

By clicking the button below, you will complete the survey and will be directed back to Mingel.

D Preregistration

Preference for Homogeneity? Occupational Segregation and Gender Stereotypes (Field Part)

Last registered on June 16, 2021

Pre-trial Fields

Trial Information

General Information

Title

Preference for Homogeneity? Occupational Segregation and Gender Stereotypes (Field Part)

RCT ID

AEARCTR-0007456

Initial registration date

May 28, 2021

Last updated

June 16, 2021 3:01 AM EDT

Location(s)

Country

Germany

Region

Primary Investigator

Name

Robert Stüber

Affiliation

WZB Berlin Social Science Center

Email

robert.stueber@wzb.eu

Other Primary Investigator(s)

PI Name

Dorothea Kübler

PI Affiliation

WZB Berlin

PI Email

dorothea.kuebler@wzb.eu

PI Name

Urs Fischbacher

PI Affiliation

University of Konstanz

PI Email

urs.fischbacher@uni-konstanz.de

Additional Trial Information

Status

In development

Start date

2021-05-31

End date

2023-08-31

Keywords

Behavior, Gender, Labor

Additional Keywords

Gender segregation, hiring decisions, teams, discrimination, stereotypes

JEL code(s)

C91; D01; J16; J71

Secondary IDs

Abstract

Many occupations are highly segregated with respect to gender. In predominantly male or female professions, there is often discrimination against the underrepresented gender. This perpetuates gender imbalances. The discrimination could be due to perceived job-specific
productivity differences between men and women. Alternatively, it could result from the belief that teams whose members have the same gender perform better. We investigate the two possible explanations in a novel experiment that varies the initial gender composition of the teams. We employ three tasks that differ with respect to gender stereotypes.

External Link(s)

Registration Citation

Citation

Fischbacher, Urs, Dorothea Kübler and Robert Stüber. 2021. "Preference for Homogeneity? Occupational Segregation and Gender Stereotypes (Field Part)." AEA RCT Registry. June 16. https://doi.org/10.1257/rct.7456-1.1.

Sponsors & Partners

Experimental Details

Interventions

Intervention(s)

- We vary the initial gender composition of the four-person group (all female or all male) and whether a woman or man is added to the group
- We also vary the task the groups perform (building shelves, memory, network).

Intervention Start Date

2021-05-31

Intervention End Date

2021-08-31

Primary Outcomes

Primary Outcomes (end points)

- Choices (binary productivity estimates) between a gender-homogeneous team that remains homogenous and a team that becomes heterogeneous (six choices, one for each combination of task (building shelves, memory, or network) and initial gender composition (all-female or all-male)).

Primary Outcomes (explanation)

Secondary Outcomes

Secondary Outcomes (end points)

- Beliefs about the choices of (eleven) other participants between a gender-homogeneous team that remains homogenous and a team that becomes heterogeneous (six choices, one for each combination of task and initial gender composition). - Binarized version of this belief measure (six choices, one for each combination of task and initial gender composition). - Choices (binary productivity beliefs) between an all-female team and an all-male team (three choices, one for each task).

Secondary Outcomes (explanation)

Experimental Design

Experimental Design

Gender-homogeneous groups consisting of four women or four men perform one of three different tasks. We then replace one worker with a new subject (either female or male), and study how this affects performance expectations and beliefs about the performance expectations of others.

Experimental Design Details

The experiment consists of three parts and a post-experimental questionnaire. In part 1, we investigate the performance assessments of groups that initially consist of either four men or four women (stage 1) where we replace one person by either a man or a woman (stage 2), that is, we elicit the productivity estimates about the performance of the new (stage 2) group that can be either more gender-heterogeneous or as homogeneous as the old (stage 1) group. We employ stereotypically male and female tasks: assembling a bookshelf and solving a memory game, and a task for which men and women are

expected to provide complementary inputs: writing down chains of movies and actors. We incentivize participants' choices by comparing their predictions to the reference performances obtained in so-called group sessions. In the group sessions, different subjects worked in groups on the tasks. In part 2, we elicit participants' beliefs about the fraction of other participants betting on each group. The participants are asked how many of 11 randomly chosen other participants bet that the group with a man performed better. We elicit subjects' beliefs about others' performance expectations for each combination of the initial group composition and task. In part 3, we also obtain a measure of the task stereotypes by eliciting participants' expectations of the performance of the all-female and all-male groups in stage 1. The experiment ends with a questionnaire in which we elicit participants' gender, age, job, and educational attainment. We also elicit participants' experience with recruitment decisions. Finally, to obtain a measure of whether the participants paid attention to their choices or not, we ask them to rate how attentive they were during the experiment.

Randomization Method

The randomization is done by a computer.

Randomization Unit

The randomization unit is the individual.

Was the treatment clustered?

No

Experiment Characteristics

Sample size: planned number of clusters

400 participants.

Sample size: planned number of observations

400 participants.

Sample size (or number of clusters) by treatment arms

400 participants (the treatment variation is within-subjects).

Minimum detectable effect size for main outcomes (accounting for sample design and clustering)

See pre-analysis plan.

Supporting Documents and Materials

Documents

Document Name

Draft: Preference for Homogeneity? Occupational Segregation and Gender Stereotypes (Lab Part)

Document Type

other

Document Description

This is a draft of the paper in which describe the laboratory part of our study. In the preregistered study we replicate our analyses using field data , i.e. using a sample of human resource and personal managers.

File

Draft: Preference for Homogeneity? Occupational Segregation and Gender Stereotypes (Lab Part)

MD5: 8ceb361c823120f230d6fa2b5405be1d

SHA1: 0e9f518b0134ac8e41c04203276d65247fa7b380

Uploaded At: April 02, 2021

Document Name

Cover letter/Email (in German)

Document Type

other

Document Description

This is the email that we will use to invite the participants (in German).

File

Cover letter/Email (in German)

MD5: 5cc8a140baaff39b6c6cccd41bc1798c

SHA1: 53c0716eb82fc941b24679b2d3f2fc41bb03d9e5

Uploaded At: May 28, 2021

Document Name

Instructions

Document Type

survey_instrument

Document Description

This document contains the experimental instructions.

File

Instructions

MD5: 8db4ec53f4576572e4ecf5c0d424529b

SHA1: fc38a306bb2674d2428ece5476255c50cc21fb18

Uploaded At: May 28, 2021

IRB

INSTITUTIONAL REVIEW BOARDS (IRBs)

IRB Name

WZB Research Ethics Committee

IRB Approval Date

2021-05-27

IRB Approval Number

2021/02/114

Analysis Plan

Analysis Plan Documents

Pre-analysis plan

MD5: 0cec699c278fdea2bfe56969ce0d063a

SHA1: 3d14d7b78f5fb363164040e91b05a74e13a1dab2

Uploaded At: May 28, 2021

Pre-analysis plan addendum

MD5: 436cbde04e75f32fa52e550ddff36ad0

SHA1: d5cc5f3ca974c053b25a75a2ae919f039e877b98

Uploaded At: June 16, 2021

Post-trial Fields

Post-trial Information

Study Withdrawal

This trial has not been withdrawn.

Intervention

Is the intervention completed?

No

Is data collection complete?

Data Publication

Data Publication

Is public data available?

No

Is there a restricted access data set available on request?

Program Files

Program Files

Reports and Papers

Relevant Paper(s)

REPORTS & OTHER MATERIALS

Pre-Analysis Plan

1. General Remarks

In this pre-analysis plan we present the analyses we plan to conduct for an online experiment. We want to analyze how personnel managers evaluate the performance of gender-homogeneous and gender-heterogeneous groups. We have already conducted this experiment with student samples in two locations (Berlin and Konstanz, Germany). Our main findings are the following. First, women are picked more often for our "female task," a memory game, and men for the "male task," building book shelves. Second, most students expect that gender-heterogeneous groups perform better than gender-homogeneous groups. We now want to investigate to which degree our findings are driven by the sample we use. We therefore conduct the experiment online with individuals who routinely make hiring decisions, that is personnel managers. To this end, we have collected the email addresses of firms/personnel managers who we invite to participate in our online experiment. The analyses we plan to perform closely follow the analyses we present in a draft of the paper in which describe the laboratory part of our study, which we attach to this preregistration ("Preference for Homogeneity? Occupational Segregation and Gender Stereotypes (Lab Part)"). We did not preregister the lab part of the study. However, the field-nature of the online experiment can potentially make a close replication of the online experiment with the personnel managers difficult. We therefore preregister here this part of our study, that is, the online experiment conducted with the personnel managers.

2. Main Research Questions

The research project answers the following main research questions:

Q1. How does the gender composition of groups (an initially gender homogeneous group that stays homogeneous or becomes more heterogeneous) affect performance expectations?

Q2. Do performance expectations respond to task-stereotypes and is there an interaction between the task stereotypes and the effect of the gender-composition of a group?

Q3. To what extent are differences in performance expectations between tasks and group gender compositions anticipated by others?

3. Experimental design

3.1 Overview

The experiment consists of three parts and a post-experimental questionnaire.

In part 1, we investigate the performance assessments of groups that initially consist of either four men or four women (stage 1) where we replace one person by either a man or a woman (stage 2), that is, we elicit the productivity estimates about the performance of the new (stage 2) group that can be either more gender-heterogeneous or as homogeneous as the old (stage 1) group. We employ stereotypically male and female tasks: assembling a bookshelf and solving a memory game, and a task for which men and women are expected to provide complementary inputs: writing down chains of movies and actors. We incentivize participants' choices by comparing their predictions to the reference performances obtained in so-called group sessions. In the group sessions, different subjects worked in groups on the tasks.

In part 2, we elicit participants' beliefs about the fraction of other participants betting on each group. The participants are asked how many of 11 randomly chosen other participants bet that the group with a man performed better. We elicit subjects' beliefs about others' performance expectations for each combination of the initial group composition and task.

In part 3, we also obtain a measure of the task stereotypes by eliciting participants' expectations of the performance of the all-female and all-male groups in stage 1.

The experiment ends with a questionnaire in which we elicit participants' gender, age, job, and educational attainment. We also elicit participants' experience with recruitment decisions. Finally, to obtain a measure of whether the participants paid attention to their choices or not, we ask them to rate how attentive they were during the experiment.

The experimental design resembles the design of the lab part of the experiment (see attachment).[1] The major differences are that i) we now ask participants to additionally evaluate the performance of the all-female and all-male groups in stage 1 of the group sessions, ii) we dropped the quantitative assessment of the performance differences between groups, iii) we added some questions due to conducting the experiment online and with a non-student sample.

3.2 Experimental Instructions

The full experimental instructions are attached to this preregistration.

4. Sample and Power Analysis

4.1 Overview

We want to conduct the experiment with a sample of personnel managers and individuals making hiring decisions in firms. We therefore collected the email addresses of individuals posting a job offer on the internet platform StepStone. StepStone is the biggest online job platform in Germany. The email addresses were collected between January 29 2021 and May 28 2021. We will send out invitations to participate in the experiment per email. We attach the email that we will use to invite the participants.

In our analysis, we will employ the full sample of participants who complete the online experiment.[2] We will base our analysis on all choices the participants make. As robustness checks, we plan to consider the sample consisting of only those participants who self-report that they were attentive when making their choices,[3] those who fully answered the questionnaire questions, those who indicate that they regularly participate in hiring decisions,[4] and/or indicate

[1] We adapt the instructions to an online experiment, which, beyond other changes, involved substantially shortening the instructions. The participants in the online experiments, however, have still access to the longer, original, descriptions of the tasks.

[2] We do not force participants to provide answers in the post-experimental questionnaire. By "completed the experiment" we mean all participants that made a choice in all 15 decisions.

[3] We plan to split the sample at the median with respect to how attentive participants were during the experiment and analyze the half that was more attentive. In contrast, we do not administer an attention check or screener, because the participants are not part of a regular panel.

[4] We plan to split the sample at the median with respect to how regularly people make hiring decisions and analyze the half that makes hiring decisions more regularly as well as to analyze the effects for those who have participated at least "occasionally."

that their occupation is in hiring, and we will also once only focus on the first choices that participants make.

4.2 Power analysis

In this power analysis, we focus on the two main results we obtained with the student sample as summarized by Result 1 and Result 2 (see also attached draft of paper):[5]

Result 1:

Most evaluators select a man for the stereotypically male task (building shelves) and a woman for the stereotypically female task (memory). In the network task, where we have no prediction with respect to stereotypes, women are chosen more often than men. The difference in choices between the tasks is reflected in the beliefs about the choices of others.

Result 2:

Most evaluators predict heterogeneous groups perform better than homogeneous groups. The degree to which others value heterogeneity is significantly underestimated. In particular, subjects fail to fully anticipate the value that others place on introducing a woman to an all-male group.

As specified in the draft, for the sake of expositional clarity, we sometimes pool several choices of one individual (see also page 11 of the draft).[6] This approach does not take into account that we consider a within-subjects design. However, all reported results are robust to using tests that take the dependence of choices into account. Table 1 reports required sample sizes for both statistical approaches we use for a power of 0.8 and a significance level of 0.05 for all findings reported as part of Result 1 and Result 2.

A sample size of about 148 fulfills these criteria. The analysis shows that a sample size of at least 148 is large enough to have sufficient power for our main analyses reported as part of Result 1 and Result 2 when we assume that we will observe the same effect sizes with the sample of personnel managers. However, because the experiment is run online, we expect substantially more noise in this sample. In addition, when deciding about the appropriate sample size, we trade off two risks: First, the risk of obtaining a sample that is overpowered, i.e., too large, such that we will be able to detect treatment effects that are statistically significant even if the treatment effect sizes we observe with the new sample are substantially smaller than in the student sample. Second, the risk of obtaining a sample that is underpowered such that the treatment differences do not turn out to be statistically significant in the new sample with the professionals. We take here what we consider to be a conservative approach and decide to err on the side of having a too large sample. The reason is that we still can interpret the effect sizes we observe and document whether the effects are smaller in the sample with personnel managers. At the same time, a large(r) sample reduces the risk of observing "too few" statistically significant treatment differences, and thereby in essence the risk of observing differences between the student sample and the sample of personnel managers – differences that we are interested in. In addition, a larger sample size allows us to investigate whether we observe heterogeneous effects depending on, for instance, participants gender or whether they regularly participate in hiring decisions. Finally,

[5] Note that this also implies that we might be underpowered to, for instance, re-detect a statistically significant difference between female and male participants.

[6] For instance, we calculate the share betting on a man for the memory and the shelves task for which individuals make two decisions each, one for the initially all-female and one for the initially all-male group.

there also is a tendency to obtain samples for online experiments that are substantially larger than for laboratory experiments which we acknowledge. We therefore plan to obtain a sample of 400 observations.

Table 1: Power analysis[7]

	Result	Test
Result 1	Most evaluators select a man for the stereotypically male task (building shelves) and a woman for the stereotypically female task (memory).	Pooled: two-sided McNemar's test (N=38), two-sided paired t-test (dependence)(N=41); by group composition: two-sided McNemar test (N=32, N =46).
Result 1	The difference in choices between the tasks is reflected in the beliefs about the choices of others.	Continuous belief outcome pooled: two-sided paired t-test (N=9), two-sided paired t-test (dependence) (N=5); continuous belief outcome by group composition: two-sided paired t-tests (N=7, N=10); binarized belief outcome pooled (N=22): by group composition: two-sided McNemar's tests (N=18, N=29).
Result 2	Most evaluators predict heterogeneous groups to perform better than homogeneous groups.	Two-sided binomial test (N=78), two-sided paired t-test (dependence)(N=12); effect of group composition by task: two-sided McNemar's exact tests (N=86, N=15, N=148)
Result 2	The degree to which others value heterogeneity is significantly underestimated (for memory and network).	Continuous belief outcome pooled: two-sided paired t-test (N=30).

4.3 Sample Generation

We collected 2743 addresses.[8] We will continue to invite individuals to participate in our experiment until our targeted sample size of 400 observations will be reached. To avoid that a single individual can participate multiple times and in order to be able to match individuals and postings, we will send out individualized links. As we do not know how high the response rate will be, we send out invitations sequentially. We will stop once our target sample size is reached.

In part 2 of the experiment, the participants are asked to assess how many of 11 randomly chosen other participants bet on a certain group. In order to be able to incentivize these choices, we first

[7] Result 1 also contains the finding that in the network task women were chosen more often than men. However, we did not have a prediction with respect to stereotypes for this task. Result 2 also contains the result that subjects failed to fully anticipate the value that others place on introducing a woman to an all-male group. Again, we initially had no hypothesis regarding this finding.

[8] Firms post continuously post new vacant positions on the platform. However, many firms use the platform for many positions/multiple times. Therefore, as the time proceeded, we collected more and more duplicates.

need to obtain these observations as a reference. We therefore preregister to obtain the first about 20 observations without belief elicitation. We will use these observations in the analysis.

4.4. Gender Balance

In the lab part of our study, we ensured that the lab sessions were gender balanced. This facilitates comparisons between female and male evaluators. We attempt to conduct the field part with a gender-balanced sample as well. Note that we do not know the gender of the participants ex-ante. However, we collected the first and surnames along with the email addresses of the people that posted a vacancy. We therefore try to infer the gender of a participant for two-thirds of the sample (67%). Out of these, about two thirds (a total of about 43%) are female and about one third (a total of 24%) are male.

In order to maximize our chances of reaching our desired sample size and in order to avoid any selection that we might introduce by primarily inviting participants from which we have the name information, we will engage in stratified sampling with the primary goal of reaching a sample that is gender balanced, and the secondary goal of reaching a sample that reflects the shares of individuals for which the first names were initially known and unknown. We will only deviate from this procedure if we cannot reach our desired sample size with this procedure. In a robustness check, we will reweight the data such that a weight of 2/3^{rds} is given to observations obtained from participants from which we initially knew their first name out of which 50% are female, and a weight of 1/3rd is given to observations obtained from participants from which we initially did not know their first name out of which 50% are female.

5. Analysis

5.1 Overview

Below, we detail the analyses we desire to conduct. Note, however, that we plan to conduct exactly the same analyses that we conducted with the student sample (see the attached draft of the paper). The only difference is that we plan to conduct some additional analyses comparing the student sample with the professional sample and that we also perform additional analyses due to conducting the experiment with the professionals online (see “Additional Analysis Professional Sample”).

5.2 Planned Analyses

Main Analysis

1. Gender and expected performance: Tasks
    - Two-sided McNemar’s test comparing the share of subjects choosing a man between the memory task and the shelves task pooling the choices for the two initial group compositions (all-female and all-male). Two-sided binomial test to analyze whether subjects prefer to add a woman or man to the group for the network task. Two-sided McNemar’s tests comparing the share of subjects choosing a man between the memory task and the shelves task separately for initially all-male or all-female groups. Test whether the difference in choices between the memory and shelves task varies by initial group composition using a difference in differences regression.

- Two-sided *t*-tests comparing the average belief about the share of subjects choosing a man between the memory task and the shelves task pooling the choices for the two initial group compositions (all-female and all-male). Two-sided *t*-tests comparing the share of subjects choosing a man between the memory task and the shelves task separately for initially all-male or all-female groups. Analogous analyses using binarized beliefs. Test whether the difference between choices and beliefs varies between the memory and shelves task using a difference in differences regression.

2. Gender and expected performance: Group composition

    - Two-sided McNemar's tests comparing the share of subjects choosing a man between the initially all-female and the initially all-male group for each task and pooled for all three tasks. Two-sided binomial test testing whether the share of heterogeneous choices differs from 50% pooling the choices for three tasks.
    - Test whether the difference in the share choosing a man between the initially all-female and the initially all-male group is significantly more pronounced for the network task in comparison to the memory or shelves task using a difference in differences regression.
    - Two-sided *t*-tests comparing the average belief about the share of subjects choosing a man between the initially all-female and the initially all-male for each task and pooled for all three tasks.
    - Test whether the average belief about the difference in the share choosing a man between the initially all-female and the initially all-male group is significantly more pronounced for the network task in comparison to the memory or shelves task using a difference in differences regression. Analogous analyses using binarized beliefs.
    - Two-sided *t*-test whether evaluators underestimate other evaluators' preference for heterogeneity. Two-sided *t*-test comparing the share of subjects correctly predict the comparative static effect of moving from an initially male to an initially female group and the task stereotype (pooled and separately by initially all-male and all-female group). Analogous analyses using binarized beliefs.

Additional Analysis

3. Gender and expected performance: Male and female evaluators

    - Two-sided McNemar's test comparing the share of subjects choosing a man between the memory task and the shelves task by group composition (all-female and all-male) and pooling the choices for the two initial group compositions by gender of the evaluator. Test whether the effect of the task stereotype varies by the gender of the evaluator using a difference in differences regression. Analogous analyses for the beliefs about the choices of others.
    - Two-sided McNemar's tests comparing the share of subjects choosing a man between the initially all-female and the initially all-male group for each task and

pooled across tasks by evaluator's gender. Binomial-tests whether the share of heterogeneous choices differs from 50% pooling the choices for the three tasks by gender. Test whether the effect of the task stereotype varies by the gender of the evaluator using a difference in differences regression. Analogous analyses for the beliefs about the choices of others.

- Two-sided McNemar's test comparing the share of male evaluators betting on a man with the share of female evaluators betting on a man while pooling across tasks and group compositions. Analogous analyses for each of the six tasks initial group composition combinations and for the beliefs about the bets of others.

4. Robustness and taking stock
    - Linear probability model of a subject choosing a man (instead of a woman) on a dummy of whether the initial group consisted of women, on the task, and on the interactions between an initially all-female group and tasks using an OLS regression (specification (1)). Specification (2) adds choice-order fixed effects and specification (3) adds individual-level control variables (age, gender, occupation).[9] We repeat these analyses using a Probit regressions.
    - Classification of subjects according to their choices exploiting our within-subjects design. We do this based on their choices for the stereotypically female and male task, shelves and memory.

Additional Analysis with the New Sample

5. To test the hypotheses that the shelves task (memory task) is stereotypically male (female), we analyze whether the share of participants who estimate that the all-male (all-female) group performs better in part 1 of the experiment is significantly different from 50% using two-sided binomial tests. We also test whether there is a difference in the share of participants who estimate that the all-male (all-female) group performs better in part 1.
6. To investigate whether there are differences in the effect of the task and/or of the initial group composition between the student and the professional sample, we conduct the regressions described under Analysis 4 and interact the task or the initial group composition, respectively, with the sample (student or professional).
7. The potential heterogeneity of the professional sample in terms of personal characteristics allows us to analyze whether the effects identified as part of the main analysis vary by gender, age, educational attainment, or occupation. We perform the regressions described under Analysis 4 and analyze their interaction with the task and the initial group composition.
8. Depending on our results, we will also consider to conduct an exploratory analysis in which we match firm and industry information to the choices we observe.

[9] In contrast to the lab sample, we also observe the educational attainment, which we will use as a control variable as well.

Pre-Analysis Plan Addendum

06/15/2021

As described in the pre-analysis plan, we collected the email addresses of firms/personnel managers who we invited to participate in our online experiment. However, the response rate was very low: out of 255 invitation that we sent out, only four participants completed our experiment. To be able to reach our targeted sample size, we now decided to collaborate with the data collecting agency *Respondi* (respondi.com) and collect a sample of personnel and general managers who regularly make hiring decisions through their panel.