Breaking ties: experimental evidence on subtle discrimination

Yee Wah Lisa Chan, Birendra Rai and Liang Choon Wang[1]

Abstract

Although legal prohibitions have reduced overt discrimination, subtle discrimination remains pervasive and difficult to detect. Using an experiment, we investigate subtle gender discrimination from systematically biased tie-breaking by employers facing equally qualified candidates. We observe subtle discrimination across different tasks, consistent with gender stereotypes. Male participants subtly discriminate against women in math tasks, while both male and female participants discriminate against men in verbal tasks. The biased tie-breaking seems inconsistent with both animus towards the out-group and standard statistical discrimination. Instead, it seems more consistent with weak in-group bias and possibly with beliefs about how others may make choices.

Keywords: discrimination, gender, beliefs, experiment

JEL codes: J16, J71, C92

[1] All authors are from Monash University. We thank audiences at Monash University, University of Groningen, University of Amsterdam (CREED), 2024 ESA-European Meeting, WZB Berlin Social Science Center, University of Konstanz (Thurgau Institute of economics), University of Exeter, Tilburg University, University of East Anglia, University of Gothenburg, University Lumière Lyon 2 (GATE-LAB) and University of Melbourne for helpful comments.

1. Introduction

Discrimination based on group characteristics such as gender and race has been illegal in most countries across the world for several decades.[2] While such laws have ostensibly reduced overt discrimination, subtle forms of discrimination that are difficult to detect remain pervasive (Foschi et al., 1994; Dovidio and Gaertner, 2000; Clermont and Schwab, 2009; Gawronski, 2019; Hebl et al., 2020; Pikulina and Ferreira, 2023). While clear evidence of subtle discrimination is scant, it is suspected to occur in numerous domains, such as racial bias in responses to rental inquiries (Christensen et al., 2021), and gender bias in the provision of career mentoring and feedback (Moss-Racusin et al., 2012) and attribution of credit in teamwork (Heilman and Haynes, 2005).

Existing studies suggest that in settings characterized by competition among participants on one side and capacity constraints on participants on the other side – such as in job hiring and promotions – subtle discrimination can occur in two main ways. First, seemingly non-discriminatory selection and promotion criteria may nevertheless systematically disadvantage certain groups (Norton et. al. 2004; Uhlmann and Cohen 2005; Barron et. al. 2024). Second, ties between equally qualified candidates belonging to different groups may (un)consciously be resolved by the decision-makers in a biased manner (Foschi, 1994; Pikulina and Ferreira 2023).

This paper develops a novel hiring experiment to detect subtle *gender* discrimination arising from systematic biases in tie-breaking, and attempts to classify whether any observed discrimination is taste-based or statistical. We focus on subtle discrimination as an aggregate

[2] For example, in the United States, the *Civil Rights Act 1964*, the *Fair Housing Act 1968*, and the *Equal Credit Opportunity Act 1974* respectively prohibit discrimination of the basis of individual characteristics, such as race and gender in labor, housing, and credit markets.

phenomenon. If an individual employer breaks ties against or in favor of woman relative to an equally qualified man, this is not clear evidence of discrimination as the employer may plausibly justify their hiring decision on some non-discriminatory grounds. But if a large number of employers *systematically* break ties in a biased manner, discrimination would be more difficult to deny.

We recruit four groups of US-based participants in our online experiment: males with college degree (MC), males with no college degree (MN), females with college degree (FC), and females with no college degree (FN). In the first stage, we collect information on the performance of the four groups. All participants answer ten incentivized questions across three categories (real-effort, verbal and math), and are not informed about their own or others' performance. Then, participants are randomly assigned to one of three treatments: the *choice* treatment, the *ordinal belief* treatment, and the *cardinal belief* treatment.[3] Participants act in the role of "employer" in the choice treatment, and act in the role of "evaluator" in the two belief treatments.

In the choice treatment, for each question, an employer had to indicate their strict *choice-ranking* over the four groups (MC, MN, FC, and FN). The higher a group is ranked, the higher the probability that a worker from that group is selected for the employer's payment (Levati et al., 2014; Cason et al., 2020). We define the *observed hiring rate* of a group as the fraction of employers who rank that group as the first-ranked group in their choice-ranking. In the ordinal belief treatment, evaluators are asked to provide their strict or weak *ordinal belief-ranking* about the average performance of the four groups for each of the ten questions in the first stage.

[3] Following Hedegaard and Tyran (2018), we use a between-subjects design to minimize confounds due to behavioral biases that may arise in a within-subjects design where the same subjects would have to provide beliefs and choices.

An evaluator earned a bonus for one randomly selected question only if their reported belief-ranking was completely correct.[4]

To detect gender bias in tie-breaking, we compare the observed gender gaps in hiring rates to believed gender gaps in performance.[5] If the gender gap in hiring rates is similar in magnitude to the believed gender gap in performance, this would imply ties are broken randomly between male and female groups. However, if we observe the gender gap in hiring rates is significantly different from the believed gender gap in performance, this would indicate ties are resolved in a biased manner – the majority of employers break their ties in favor of one gender.

We first focus on detecting hiring discrimination among the two groups with college education – the MC and FC groups – as they are ranked first by the vast majority of employers in their choice-rankings.[6] We find no significant gender differences in actual performance between the MC and FC groups in any task category. In the ordinal belief treatment, the majority of evaluators report a tie in their first rank involving the MC or FC groups. However, a non-trivial minority of evaluators believe there are significant gender differences in performance. Consequently, on average, beliefs are biased in favor of women on the verbal task and biased against women on the math task.[7] As there are no gender differences in performance, these

[4] In the cardinal belief treatment, for each question, evaluators provide their cardinal beliefs about the performance of each group, i.e., four integer numbers, each ranging from zero to 100. The four numbers reported by an evaluator for a question indicate the evaluator's beliefs about the percentage of participants in each group that correctly answered that question in the first stage. The evaluators in the cardinal belief treatment earned a bonus for their belief about each group in one randomly selected question (if their reported cardinal belief about a group was within five percentage points of the actual percentage of correct answers by participants in that group).

[5] Methodologically, random assignment of participants to treatments ensures the validity of comparing choice-rankings and belief-rankings provided by different set of participants. Balance tests suggest there are no significant differences in the characteristics of participants in the choice and belief treatments (Appendix Table A.1).

[6] Our main analysis involves comparing the choice and ordinal belief treatments. We use comparisons between the choice and cardinal belief treatments to conduct robustness checks. One reason being that we suspect it is harder for participants to form cardinal beliefs about groups than to provide an ordinal ranking over the groups.

[7] For the purposes of our study, it is secondary whether or not ties reflect a genuine belief that the tied groups have similar performance. The possibility that ties may reflect "incompleteness", i.e. an evaluator's inability to arrive at a relative comparison between groups, does not provide a reason for systematic biases in tie-breaking.

beliefs are inaccurate and stereotypical, which is consistent with previous studies (Reuben et al., 2014; Bordalo et al., 2016; Bordalo et al., 2019; Bohren et.al., 2023). However, we observe gender discrimination in hiring: employers in the choice treatment are significantly less likely to hire men relative to women in the real-effort and verbal tasks, and significantly less likely to hire women relative to men in the math task.

Our main finding is that ties are broken in a biased manner, but it depends on the domain and gender of the employer. The majority of employers break their ties in favor of women in the verbal task and against women in the math task. This is consistent with gender stereotypes that women are relatively better at verbal while men are relatively better at math (Coffman, 2014; Reuben et al., 2014; Bordalo et al., 2019). Although all employers discriminate in favor of women in the verbal task, male employers discriminate against women in the math task and female employers discriminate in favor of women in the real-effort task.

Our findings are largely robust. We show that the propensity to state ties in beliefs is unlikely to be driven by social desirability bias. Most of the results also hold when we focus on men and women *without* college degrees. One exception arises in the math task where we do not detect subtle discrimination among both male and female employers. Perhaps participants in our study view math as a domain that requires college education. This result is important because it highlights the possibility of unbiased tie-breaking. As such, it adds to the internal validity of our findings about the existence of subtle discrimination in gender comparisons among participants with college degree where ties are broken in a biased manner.[8]

[8] With some minor caveats, all these findings continue to hold when we compare the choice treatment with the cardinal belief treatment. For this comparison, we first construct the *implied* ordinal belief-ranking from an evaluator's reported cardinal beliefs about the performance of each group. Subsequently, we use the same method of analysis as used in comparing the choice treatment with the ordinal belief treatment.

We argue the observed subtle discrimination in our study may be driven by both preferences and beliefs. However, the biased tie-breaking seems inconsistent with animus towards the out-group and standard statistical discrimination. Instead, it seems more consistent with weak in-group bias and possibly with beliefs about how others would make choices.

Our finding that subtle discrimination via biased tie-breaking arises in every task category complements the existing literature on discrimination in a novel way. In correspondence studies which are regarded as the "gold standard" for empirical investigation of discrimination (Riach and Rich, 2002; Bertrand and Mullainathan, 2004; Bertrand and Duflo, 2017; Gaddis, 2018), potential employers have to choose between fictitious candidates that are identical except for their group identity. These studies do not elicit employers' beliefs. Observed differences in interview call-back rates across candidates of different group identities are typically attributed to employers believing the average productivity of groups are *different*. Existing experimental studies have also focused on showing that it is the differences in beliefs about average performance of groups – whether accurate or inaccurate – which drives discrimination (Reuben et al., 2014; Coffman et al., 2021; Bohren et al., 2023; Barron et. al. 2024), and on identifying implicit preferences or discrimination from choices (Barron et. al. 2024; Cunningham and de Quidt 2024). In contrast, we link employers' beliefs and choices and focus on how ties are resolved, showing that biased tie-breaking creates gender differences in hiring rates even among employers who believe the performance of equally qualified men and women are *similar*.

The rest of the paper is organized as follows. Section 2 describes our experimental design. Section 3 describes our primary outcome measures and statistical tests. Section 4 reports the results. Section 5 presents the robustness checks. Section 6 concludes.

2. Experimental design

The online experiment was conducted in May 2024. A total of 451 participants based in the U.S. aged 25-64 years were recruited via Prolific, a UK-based online research platform. The sample was stratified to create a demographically representative sample in terms of age, gender and education (college degree) using data from the American Community Survey 2022. We recruited four groups of participants: MC, FC, MN and FN.

The experiment had three main stages (Table 1). In Stage 1, all participants completed the individual task. In Stage 2, each participant assessed their own performance in each category. In Stage 3, we randomly assigned participants to one of three treatments, where they made hiring decisions (choice treatment) or evaluations (ordinal belief or cardinal belief treatments).

Table 1 – Summary of stages in the experiment

Stage	Description
0	Consent
1	Timed individual task involving 10 questions (2 real-effort questions, 4 verbal questions, and 4 math questions)
2	Beliefs about own performance
3	Random assignment to treatments: choice, ordinal belief or cardinal belief • Choice treatment: elicitation of choice-rankings over the four groups • Ordinal belief treatment: elicitation of ordinal belief-rankings about performance of the four groups • Cardinal belief treatment: elicitation of cardinal beliefs about performance of the four groups
4	Demographic survey

Before each stage, participants completed incentivized comprehension questions about the instructions and received feedback on the correct answer to these questions. Following completion of Stage 3, participants answered demographic questions about themselves. On average, participants took about 19 minutes to complete the experiment and earned GBP 3.00.

2.1 *Details of the experiment*

We describe the experimental design in detail below.

Stage 1: Individual task – Participants answered ten questions in three categories: real effort, verbal and math. We selected these tasks because they require both skill and effort, and because previous studies suggest there are no gender gaps in performance (Niederle and Vesterlund 2007; Bordalo et al., 2019; Bohren et al., 2023).[9] Participants had a maximum of thirty seconds per question. They earned GBP 0.10 per correct answer, lost GBP 0.02 per incorrect answer, but were not penalized for skipping questions. Questions were presented on a separate page and in the same order for all participants. We collected data on performance for each of the four groups.

Stage 2: Beliefs about own performance – Participants were asked to indicate, for each question in Stage 1, whether they believed they got the question correct. They earned GBP 0.05 per correct guess. Questions were presented on the same page by the order of questions in Stage 1.

Stage 3: Treatment Interventions – Participants are randomly assigned to one of three treatments in a between-subject design. This allows us to reduce potential confounds from behavioral biases which may lead to measurement error in the extent of observed bias in beliefs and discrimination (Hedegaard and Tyran, 2018). For example, participants may seek consistency between their belief and choice-rankings: a decision-maker may first discriminate against women and then report beliefs of gender gaps in performance to justify this decision (Bohren et. al., 2023). Therefore, we elicit employers' choices and beliefs from separate groups of participants.

[9] The real effort task involved counting the number of zeros in a sequence of numbers, and decoding numbers to letters. The verbal task required rearranging letters to form a new word (anagrams) and finding a smaller sub-word within a larger word. The math task involved adding and dividing numbers.

In all treatments, participants revisit the ten questions seen in Stage 1. They make ten decisions corresponding to each question and are paid for one randomly selected decision. Participants receive no feedback on the performance of the four groups. Decisions were presented on separate pages by the order of questions presented in the individual task.[10]

Choice treatment – Participants act as employers and are asked to provide strict choice-rankings over each of the four groups. Ties are not allowed to mimic settings where only one candidate can be hired. Employers were paid based on the performance of a randomly selected worker, earning GBP 1.00 if the worker answered the question correctly. The higher a group is ranked, the higher the probability an individual would be selected from that group for the employer's payment.[11] This provides incentives for employers to rank groups in decreasing order of believed average group performance (Levati et al., 2014; Cason et al., 2020). The selected worker was not paid to rule out employers' other-regarding preferences with respect to workers.

Ordinal Belief Treatment – Participants act as evaluators and are asked to provide their ordinal belief-rankings about the performance of the four groups. As we wish to examine subtle discrimination from tie-breaking, ties were allowed in the belief-rankings. A participant earned GBP 1.00 only if their belief-ranking in one randomly selected question was completely correct.

[10] To account for possible order effects, the questions in all three treatments were randomly presented with either (Male College and Male No College) shown first, or (Female College and Female No College) shown first.

[11] The most preferred group had a 60% chance of being selected, the second most preferred group had a 30% chance of being selected, the third most preferred group had a 10% chance of being selected, and the least preferred group had 0% chance of being selected.

Cardinal Belief Treatment – Participants also act as evaluators but are asked to provide their cardinal beliefs about the performance of the four groups. For each question, they are asked to guess the percentage of participants in each of the four groups who answered the question correctly. Participants earned GBP 0.25 per reasonably accurate guess (within five percentage points of the true percentage) in one randomly selected question.

3. Primary outcomes and empirical approach

To test for subtle discrimination arising from biased tie-breaking, we focus on comparing the choice and ordinal belief treatments, and use the cardinal belief treatment to check robustness (Section 5). For our analysis, we first construct several outcome measures at the individual-level using data on an individual's performance and choice or belief-rankings. All measures are constructed separately for each of the four groups and for each task category. We use the individual-level outcomes to construct group-level outcomes. We describe the construction of these variables in more detail below.

Performance score of a group – For each participant, we construct a measure of performance in a category by calculating the average number of questions answered correctly over all questions in the task category. As each individual belongs to one of the four groups, the performance scores of all individuals in a group are then used to construct the distribution of performance scores of a group.[12] We use the distributions of performance scores in any pair of male and female groups to check for gender gaps in actual performance.

[12] In our analysis, we pool the data in all three treatments to construct performance distributions for each group as participants answered the ten questions prior to being assigned to treatments, and did not have any information about the subsequent stages of the experiment.

Observed hiring rate of a group and hiring gap between groups in a task category – In the choice treatment, we assume the first-ranked group in the strict choice-ranking represents the group hired by the employer. For each individual employer, the hiring rate of a group in a task category is defined as the fraction of questions in the category where the group is ranked first. Each employer's hiring rate of a group in a task category is used to construct the corresponding distribution of hiring rates of a group for that category. For each individual employer, we also construct the difference in the hiring rates between male and female groups in a task category (i.e. MC – FC, MN – FN). This gives us, for each employer, a measure of the observed gender gap in hiring in a task category, allowing us to construct the distribution of the observed hiring gap. We use the observed hiring gap to check for gender discrimination in hiring.

Beliefs about performance of a group in a task category – In the ordinal belief treatment, we assume the first-ranked group represents the group most likely to be hired, and focus on the first rank because it is most comparable to the first rank in the choice treatment. Recall in the ordinal belief treatment, evaluators can provide weak or strict rankings over the four groups. Thus, one or more groups may be ranked first in a belief-ranking. For each individual evaluator, we calculate the fraction of questions in a category where a particular group is ranked first using two measures: (i) *ties* at first rank (e.g. both MC and FC groups are ranked first), and (ii) *unique* first rank (e.g. MC group is ranked first but FC group is not, and vice versa). Gender differences in propensity for a group to be uniquely ranked first represents believed gender gaps in performance.

To infer gender bias in tie-breaking, we compare the observed gender gaps in hiring rates to believed gender gaps in performance. As employers and evaluators are observationally similar (Appendix Table A.1), we interpret the comparison between choices and beliefs as if they are from the same group of participants (i.e., employers). If the gender gap in hiring rates is similar in magnitude to the believed gender gap in performance, this would imply ties are broken randomly between the MC and FC groups. However, if we observe the gender gap in hiring rates to be significantly different from the believed gender gap in performance, this would indicate ties are resolved in a biased manner.

We begin by examining gender gaps in performance in Section 4.1, showing that there are no significant gender differences in performance in all task categories. We examine evaluators' beliefs in Section 4.2, and document that the majority of evaluators state ties in the believed average performance of men and women, while a minority of evaluators believe there are gender gaps in performance in line with gender stereotypes. We consider employers' hiring decisions in Section 4.3, showing how observed hiring discrimination varies by task category and the gender of employers. Section 4.4 compares beliefs to choices, revealing that there is bias in how ties are resolved. We discuss the robustness of our findings in Section 5. Throughout our analysis, we use the Wilcoxon-Mann Whitney rank-sum test for between-subject comparisons, and the Wilcoxon signed-rank test for within-subject comparisons. This is because none of the variables are normally distributed according to a Kolmogorov-Smirnov test.

4. Results

Before we present the main results, we report the distribution of hiring rates across the four groups. Table 2 presents employers' average hiring rates for each group in the choice treatment. MC and FC account for more than 80 per cent of groups hired in all task categories. This is true at the aggregate level and regardless of the gender of the employer. Therefore, we initially

Table 2 - Observed hiring rates of the four groups by employers in the Choice treatment

	Real effort (1)	Verbal (2)	Math (3)
A. All employers			
MC	0.363	0.283	0.540
FC	0.483	0.607	0.352
Total	0.846	0.890	0.892
MN	0.080	0.048	0.063
FN	0.073	0.062	0.045
Total	0.153	0.110	0.108
Obs.	150	150	150
B. Male employers			
MC	0.453	0.323	0.583
FC	0.413	0.573	0.307
Total	0.866	0.896	0.890
MN	0.080	0.047	0.063
FN	0.053	0.057	0.047
Total	0.133	0.103	0.110
Obs.	75	75	75
C. Female employers			
MC	0.273	0.243	0.497
FC	0.553	0.640	0.397
Total	0.827	0.883	0.893
MN	0.080	0.050	0.063
FN	0.093	0.067	0.043
Total	0.173	0.117	0.106
Obs.	75	75	75

focus on the choices and beliefs regarding MC and FC. Unless otherwise stated, gender differences in the following sections will refer to differences between the MC and FC groups.

4.1 *Gender gaps in performance*

Figure 1 reports the average performance scores by gender. Women perform slightly better than men in the real effort and verbal tasks, while men perform slightly better than women in the math task. However, rank-sum tests indicate these performance differences are not significant and that the distributions are not significantly different (p = 0.245 for real-effort; p = 0.333 for verbal; p = 0.207 for math).[13] This result is in line with previous studies showing no gender differences in performance in simple math (Niederle and Vesterlund 2007; Bordalo et al., 2019; Bohren et al., 2023) and verbal tasks (Dreber et. al. 2014). Furthermore, a variance

[13] We report all pairwise comparisons of average performance between each of the four groups in Appendix Table A.1. We also find no significant gender difference in the median performance across any task (p = 0.356 for real effort; p = 0.321 for verbal; p =0.371 for math).

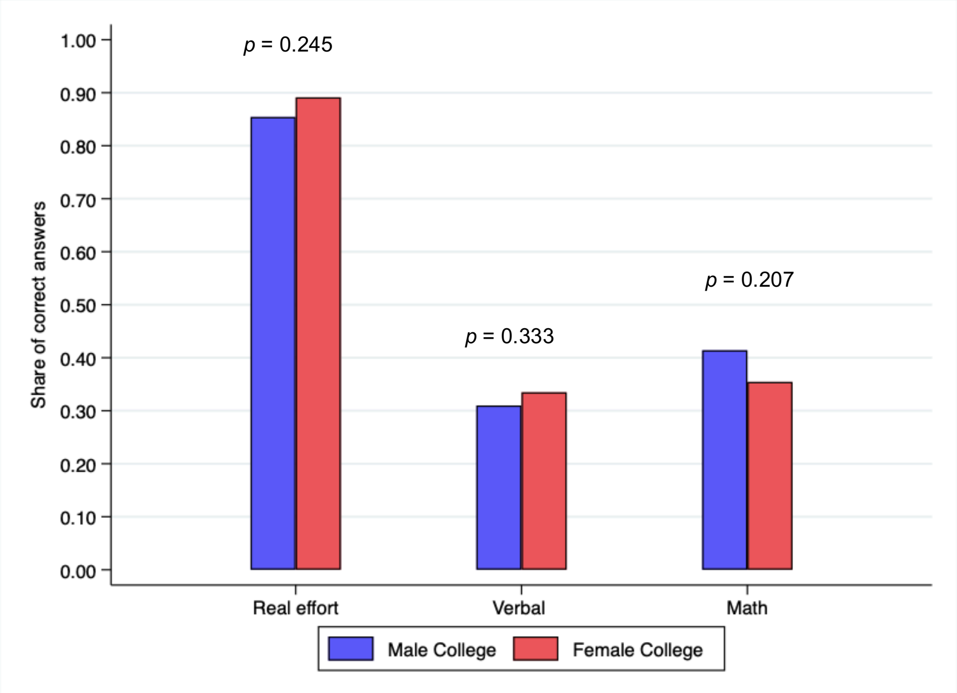

Figure 1 – Gender differences in performance between the MC and FC groups

ratio test that is robust to non-normality and uses the median (Levene 1960; Brown and Forsythe 1974) suggests there is no significant gender difference in the variance of performance scores in any task category (*p*=0.298 for real-effort; *p*=0.481 for verbal; *p*=0.106 for math).

Result 1: There is no significant gender gap in performance in any task category.

4.2 *Believed gender gaps in performance*

We have thus far shown that the actual performance of men and women is similar. We next examine whether there are any believed gender gaps in performance by checking whether women or men are more likely to be ranked first in the believed performance rankings of evaluators in the ordinal belief treatment. Recall evaluators could provide either strict or weak ordinal rankings over the four groups; therefore, more than one group could be ranked first.

Panel A in Figure 2 shows how often men and women are *both* weakly ranked first in the believed performance rankings. The majority of evaluators state ties, indicating they believe there is no gender difference in performance across all tasks. The level of ties depends on the task category: the share of ties in the real-effort task is higher than in the verbal and real-effort tasks. However, a non-trivial minority of evaluators believe there are gender differences in performance. Panel B in Figure 2 shows how often men and women are *uniquely* (i.e. strictly) ranked first in the believed performance rankings, where one gender is ranked first while the other is not. Among these minority of evaluators, they believe men are relatively better at the math task, women are relatively better at the verbal task, although they do not believe there is

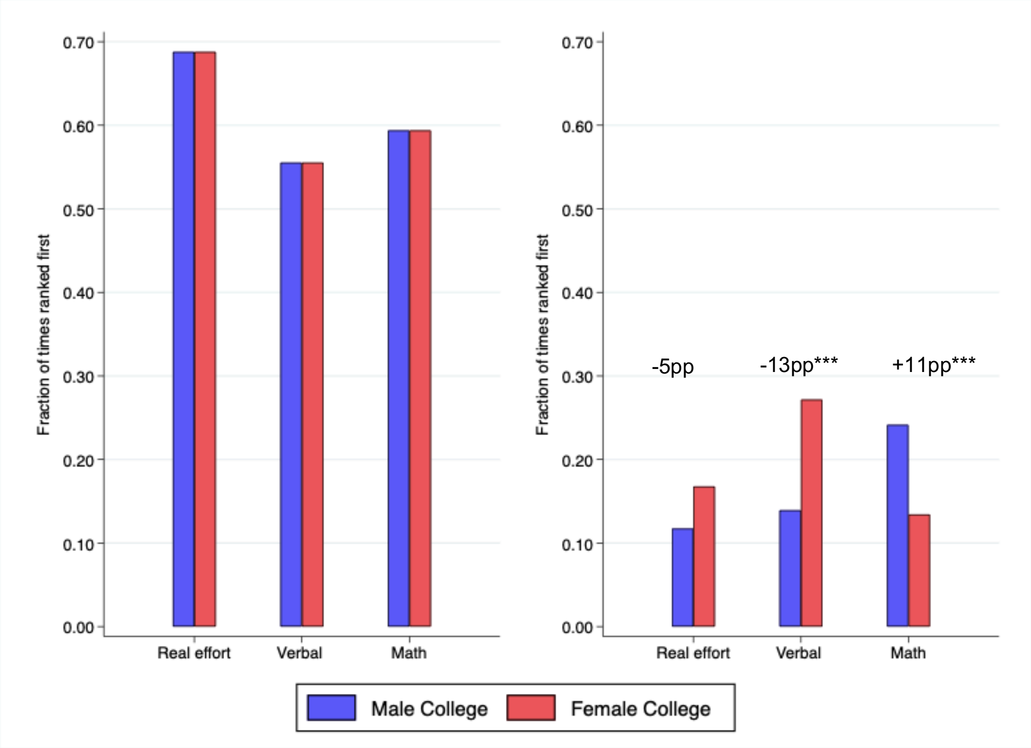

Figure 2 – Frequencies of the MC and FC groups being ranked first in performance rankings of evaluators in the Ordinal belief treatment.

Notes: Panel A shows the average fraction of times both groups are ranked first. Panel B shows the average fraction of times a group is uniquely ranked first. The reported differences refer to the difference in the average fraction of times a group is uniquely ranked first between the MC and FC groups (MC-FC). The reported *p*-values corresponding to the null hypothesis of no difference between the groups using the Wilcoxon signed-rank test are *p*=0.212 for real effort, *p*=0.000 for verbal and *p*=0.003 for math.

any gender difference in performance in the real-effort task. The believed gender difference in performance is just over 10 percentage points in favor of women in verbal and in favor of men in math. Given there are no gender differences in actual performance, these belief gaps are inaccurate and in line with gender stereotypes (Bordalo et. al. 2016; Bordalo et. al. 2019). Since the majority of belief-rankings in every task category involve ties at the first rank, any believed gender differences in performance are driven by a minority of evaluators.

Table 3 reports believed gender gaps in performance at the aggregate level, and separately for male and female evaluators in each task category. Note the believed gender gaps are based on the difference in the fraction of times the MC and FC groups are uniquely ranked first. Believed gender gaps in performance also depend on the gender of evaluators. While female evaluators believe women perform relatively better in the real-effort task, male evaluators believe men perform relatively better in the math task. In contrast, all evaluators believe women perform relatively better in the verbal task, with the size of the believed female advantage similar across male and female evaluators (12 and 14 percentage points respectively).

Table 3 – Difference in the frequencies of the MC and FC groups being uniquely ranked first in performance rankings of evaluators in the Ordinal belief treatment

	(1) Real effort	(2) Verbal	(3) Math
A. All evaluators			
Belief gap	-0.050	-0.133***	0.107***
	(0.212)	(0.000)	(0.003)
Obs.	149	149	149
B. Male evaluators			
Belief gap	0.054	-0.122**	0.162**
	(0.131)	(0.015)	(0.010)
Obs.	74	74	74
C. Female evaluators			
Belief gap	-0.153***	-0.143***	0.053
	(0.002)	(0.006)	(0.115)
Obs.	75	75	75

Notes: p-value reported in parentheses. *** p<0.01; ** p<0.05; * p<0.10

Result 2: The majority of evaluators state ties in the believed performance of men and women. A minority of evaluators state beliefs in line with gender stereotypes that men are relatively better at math and women are relatively better at verbal tasks. Beliefs about gender gaps in performance vary by the gender of evaluators.

4.3 *Hiring gaps*

We next examine whether there is differential hiring of men and women using the choice-rankings of employers in the choice treatment. A significant difference in the hiring of men relative to women would indicate gender discrimination in hiring.[14]

Recall employers were asked to strictly rank the four groups, and we interpret the first ranked group as the group from which they would hire. Figure 3 shows the average fraction of times the MC and FC groups are ranked first in each task category. Large and significant gender gaps in hiring are observed in every task category. Employers hire women significantly more often than men for real-effort and verbal tasks ($p = 0.056$ for real effort; $p < 0.001$ for verbal). In the math task, the pattern is reversed: employers are significantly less likely to hire women than men ($p = 0.002$). This is consistent with previous studies which find hiring discrimination against women in math tasks (Reuben et al., 2014; Coffman et al., 2021).

[14] We focus on direct discrimination, which is defined as a causal link between differential treatment and group identity, holding fixed all other observable characteristics (Bohren et. al. 2022; Bohren et. al. 2023). It is easier to interpret differential treatment based on group identity as discrimination when there are no observed average productivity differences between groups (Neumark 2012). Classifying the observed discrimination as taste-based or statistical is a separate matter.

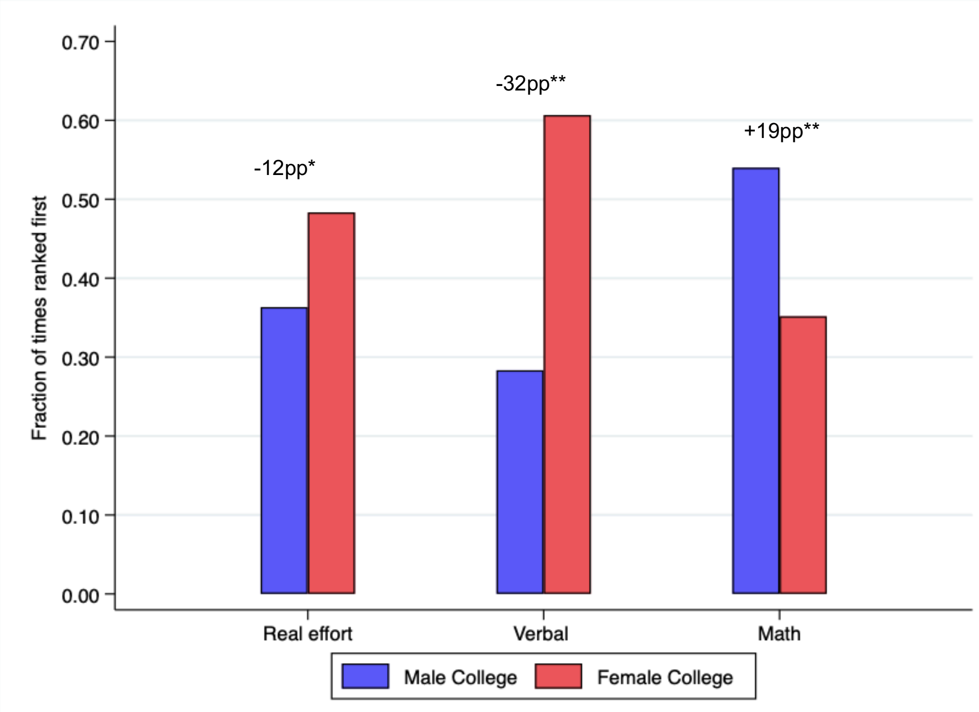

Figure 3 – Observed hiring rates of the MC and FC groups by employers in the Choice treatment

Notes: The reported *p*-values corresponding to the null hypothesis of no difference between the groups using the Wilcoxon signed-rank test are p=0.056 for real effort, p<0.001 for verbal and p=0.002 for math.

Result 3: Hiring discrimination is observed in every task category. Employers discriminate in favor of women in the real-effort and verbal tasks, and against women in the math task.

Table 4 reports the observed hiring gaps in each task category among all employers, and separately among male and female employers. The extent of hiring discrimination varies by the gender of employers across task categories.

Table 4 – Observed hiring gap between MC and FC groups in the Choice treatment

	(1) Real effort	(2) Verbal	(3) Math
A. All employers			
Hiring gap	-0.120*	-0.323***	0.188***
	(0.056)	(0.000)	(0.002)
Obs.	150	150	150
B. Male employers			
Hiring gap	0.040	-0.250***	0.277***
	(0.654)	(0.005)	(0.002)
Obs.	75	75	75
C. Female employers			
Hiring gap	-0.280***	-0.397***	0.100
	(0.002)	(0.000)	(0.226)
Obs.	75	75	75

Notes: p-values reported in parentheses correspond to the null hypothesis of no hiring gap using the Wilcoxon signed-rank test on matched pairs. *** p<0.01; ** p<0.05; * p<0.10.

Result 4: Hiring discrimination in favor of women in the real-effort task is only observed among female employers, while discrimination against women in the math task is only observed among male employers. However, discrimination in favor of women in the verbal task is observed among all employers.

4.4 *Detecting whether ties are randomly broken – comparing beliefs and choices*

We have shown that despite the majority of evaluators stating they believe there are no gender differences in performance (i.e. ties), we observe significant gender discrimination in hiring across all task categories. Notably, gender gaps in hiring appear to exceed the believed gender gap in performance of 10 percentage points in verbal and math tasks, suggesting there may be some bias in tie-breaking. One way to infer bias in tie-breaking is to check whether there exists an excess hiring gap, i.e., a significant difference between hiring gap and belief gap. A significant excess hiring gap would provide evidence of biased tie-breaking, thereby indicating subtle discrimination.

Table 5 reports the excess hiring gaps among all employers, and separately for male and female employers. Excess hiring gaps are significant in verbal and math, implying the majority of

Table 5 – Excess hiring gap between MC and FC groups across task categories

	(1) Real effort	(2) Verbal	(3) Math
A. All participants			
Excess hiring gap	-0.069	-0.191***	0.081*
	(0.199)	(0.000)	(0.055)
Obs. (Choice treatment)	150	150	150
Obs. (Ordinal belief treatment)	149	149	149
B. Male participants			
Excess hiring gap	-0.014	-0.128*	0.115*
	(0.931)	(0.061)	(0.069)
Obs. (Choice treatment)	75	75	75
Obs. (Ordinal belief treatment)	74	74	74
C. Female participants			
Excess hiring gap	-0.127*	-0.253***	0.047
	(0.063)	(0.002)	(0.479)
Obs. (Choice treatment)	75	75	75
Obs. (Ordinal belief treatment)	75	75	75

Notes: p-value reported in parentheses. *** p<0.01; ** p<0.05; * p<0.10

employers break their ties in favor of women in verbal and against women in math. Again, bias in tie-breaking seems to depend on the gender of employers, although these gender differences appear to be marginally significant. Male employers break their ties against women in math, while female employers break their ties in favor of women in the real-effort task. Although all employers break their ties in favor of women in the verbal task, the size of the bias in favor of women appears somewhat larger among female employers.

Result 5: The majority of employers break their ties in favor of women in the verbal task and against women in the math task. Subtle discrimination is driven by female employers in favor of women in the real-effort task, by all employers in favor of women in the verbal task, and by male employers against women in the math task.

If we take the beliefs of evaluators at face value (independent of whether they are accurate or not), then a hiring gap up to the size of the belief gap can be completely rationalized by theories of statistical discrimination. However, since theories of statistical discrimination are silent about how ties get resolved, it is not clear whether these excess hiring gaps should be interpreted as driven by beliefs or preferences (statistical or taste-based discrimination). We revisit this issue in the discussion and conclusion section (Section 6).

4.5 *MN versus FN groups*

We now turn to the comparisons for the two groups without a college degree, i.e. the MN and FN groups. We conduct these comparisons in a slightly different way. As shown in Table 2, the MN and FN groups are ranked first in the choice-rankings of employers far less frequently than the MC and FC groups. Hence, in comparing the MN and FN groups we focus on the *relative* ranking between MN and FN groups (i.e. whether the MN group is ranked higher than, lower than, or equal to the FN group) instead of absolute first ranking.

Most of the results obtained by comparing the MC and FC groups continue to hold in the comparison of MN and FN groups (Appendix Tables A.14 to A.18). The most significant difference is that although we still observe subtle discrimination in favor of women in the verbal task and also in the real-effort task, we do not find any evidence of subtle discrimination for or against women in the math task. Appendix Table A.16 shows there is no subtle discrimination in the math task among all employers, regardless of the gender of employer – suggesting unbiased tie-breaking between the MN and FN groups in the math task. This in contrast to our earlier finding for the MC and FC groups that ties are broken in favor of men in the math task among male employers.

Result 6: There is no evidence of subtle gender discrimination in the math task between participants without a college degree by female or male employers.

5. Robustness checks

One potential concern is social desirability bias. The idea is that evaluators may be willing to forgo earnings to appear non-discriminatory or non-sexist. For example, some evaluators might believe men are relatively better at math, but instead state a tie in their beliefs due to image concerns. To assess whether the propensity to state ties is sensitive to social desirability bias, we divide the sample into evaluators with low versus high tendency to give socially desirable answers. We construct a social desirability score based on the 13-item short form Marlowe-Crowne social desirability scale after the experiment (Reynolds 1982), which measures the propensity to give socially desirable answers by asking participants whether they have certain personality traits. Following Dhar, Jain and Jayachandran (2022), we classify evaluators based on whether their social desirability score is above or below the median score among all evaluators. We find that there are largely no significant differences in the propensity to state ties in beliefs between evaluators with low and high social desirability scores (Appendix Tables A.20 – A.21). We also find the correlation between the social desirability score and the fraction of ties in a task category is close to zero. This evidence suggest ties are unlikely to be driven by social desirability bias.

Our main results for the MC and FC groups used the beliefs and choices regarding how often they are ranked first. When we consider the relative rankings between the MC and FC groups, our findings remain similar, albeit a bit weaker (Appendix Tables A.5 to A.7). Although we do not find evidence of subtle discrimination against women in the math task at the aggregate level, we continue to find evidence of biased tie-breaking in favor of men in the math task among male employers. This reflects the intuitive idea that the extent of bias in tie-breaking between the MC and FC groups depends on how consequential the decision is for the employers. The resolution of ties at the first rank has larger implications for the payoffs of employers than the resolution of ties at lower ranks.

Finally, we conduct all the above-mentioned comparisons using the choice treatment and the *cardinal* belief treatment. We first construct the implicit ordinal belief-ranking for each cardinal belief-ranking, and follow the same empirical strategy as used in comparing the choice and ordinal belief treatments. Although ties are less common in the cardinal belief treatment than in the ordinal belief treatment, all the findings continue to hold qualitatively (Appendix Tables A.8 to A.14).

6. Discussion and conclusion

The literature on discrimination uses the terms "overt" and "subtle" discrimination to emphasize that some discriminatory behaviors are easier to detect than others. We conduct an experiment that is designed to detect subtle discrimination arising from biased tie-breaking. We find clear evidence of subtle discrimination in hiring in every task category in the direction consistent with gender stereotypes, showing that biased tie-breaking leads to hiring discrimination. This is despite the majority of evaluators stating ties in the believed performance rankings of men and women, and only a minority of evaluators reporting inaccurate and stereotypical beliefs about gender. Our novel finding regarding the role of tie-breaking illustrates that differential treatment of groups is possible even when employers believe the average performance of groups is similar.

Following the seminal work of Becker (1957), a key challenge for studies on discrimination is classifying whether discrimination is taste-based or statistical, as it is crucial for determining appropriate policies to address discrimination (Phelps, 1972; Arrow, 1973; Aigner and Cain, 1977; Bertrand and Mullaianathan, 2004; Ewens et al., 2014; Hedegaard and Tyran, 2018; Coffman et al., 2021). In addition, over the last two decades, the literature has increasingly emphasised that statistical discrimination may be driven by either accurate or inaccurate beliefs (Arrow, 1998; List, 2004; Mobius and Rosenblat, 2006; Schwartzstein, 2014; Bohren et al., 2019; Bohren et al., 2023; Barron et al., 2024; Islam et al., 2024).

In general, both taste-based and belief-based sources may contribute to biased tie-breaking in our study. For example, biased tie-breaking across task categories occurs in the direction of gender stereotypes. This suggests participants may be relying on stereotypes to resolve ties.

"Taste" can contribute to biased tie-breaking in two ways: animus towards the "out-group" or favoritism towards the "in-group". To the best of our knowledge, no prior experimental study on discrimination has found unambiguous evidence of animus towards the out-group (Bar and Zussman, 2020; Bohren et al., 2023). This is most likely because the stylized experimental interactions are temporary and anonymous; in our online experiment, "employers" know they do not have to meaningfully interact with the "workers".

However, it is possible that what appears to be animus towards the "out-group" could instead be favoritism towards the "in-group". One test of in-group bias is to examine whether employers are willing to pay a cost in order to favor their own gender – do employers favor their own gender even when they believe the other gender is clearly better? At the aggregate level, we do not observe this. Whenever we observe gender discrimination in a domain, we also observe a corresponding large belief gap in that domain. Thus, it appears employers choose to hire the gender who they believe is clearly better. To the extent that in-group bias is present in our study, it does not generalize to both genders or all tasks. If anything, in-group bias seems somewhat stronger among female employers. Male employers break their ties in favor of women in the verbal task, consistent with their beliefs that women are relatively better than men. In contrast, female employers do not break their ties against women in the math task despite believing that women perform relatively worse than men. This alternative interpretation is supported by prior studies which find some evidence of in-group favoritism among women in broadly similar contexts (Bagues and Esteve-Volari, 2010; De Paola and Scoppa 2015; Coffman et. al. 2021; Bohren et. al. 2023; Cappelen et. al. 2023).

Given that in-group bias in our study seems weak at best, we consider whether the observed biased tie-breaking is driven by beliefs. The theory of statistical discrimination typically focuses on believed differences in the *mean* performance. By design, ties suggest employers believe there is no gender difference in average performance, which therefore rules out statistical discrimination based on mean group differences.

Even if employers believe there are no mean group differences, they might still believe there are group differences in the *variance* of performance. Suppose employers are risk averse. Then the choice of men over women in the math task would suggest they believe there is some advantage in doing so, even when they believe there are no average differences between men and women. The advantage can potentially be the upside that there are more high performing men in math compared to women. If this is the advantage, then by the fact that men and women have the same mean for employers with ties, there must also be more low performing men in math such that the variance of men is larger than that of women to keep the mean the same across the two groups. If employers are risk averse, the larger variance for men implies that men should be less likely to be hired. Thus, believed gender differences in the variance of performance is unlikely to be the reason for why these employers prefer men over women in the math task.

To check whether employers do actually believe there are gender differences in the variance, we can use data on the beliefs about the mean performance score in each group in the cardinal belief treatment.[15] We find no evidence that employers believe there is a gender difference in the variance. Therefore, statistical discrimination driven by believed group differences in the variance of performance seems unlikely.

One possibility is that the observed biased tie-breaking is driven by beliefs about how others would act. Arguably, individuals are aware of the gender stereotypes in a task category that others hold and use when they make their choices. It is therefore possible that employers who have to make a choice and feel unable to decide themselves may rely on these common stereotypes to break their ties, thus making their choice in line with what they perceive many others would do.

Our findings have potentially important policy implications. We detect subtle hiring discrimination in our stark environment with two hiring stages and where participants are incentivized to make decisions which maximized their earnings. In real hiring environments, organizations may have multiple hiring stages to narrow down the pool of candidates, and decision-makers may not necessarily have any material stakes in evaluation or hiring decisions of candidates (e.g. HR personnel who evaluate CVs). Furthermore, we were able to directly observe both hiring and evaluation decisions in our study, whereas full transparency of these decisions in real organizations is unlikely for confidentiality reasons. Therefore, subtle

[15] We first calculate the mean believed performance score across all questions in a task, and then find the variance associated with that mean.

discrimination may be more difficult to detect in practice, and our findings should be taken as lower bound estimates of potential subtle discrimination in hiring. Future work should investigate whether subtle discrimination is generalizable to other settings and group identities (such a race or ethnicity), particularly in the field.

References

Aigner, D.J., and G.G. Cain (1977) "Statistical Theories of Discrimination in Labor Markets," *Industrial and Labor Relations Review*: 30(2), 175–187.

Arrow, K.J. (1973) "The Theory of Discrimination." in O. Ashenfelter and A. Rees, Princeton (Eds.), *Discrimination in Labor Markets*. New Jersey: Princeton University Press.

Arrow, K.J. (1998) "What has economics to say about racial discrimination?" *Journal of economic perspectives*, 12(2): 91-100.

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

Bar, R., and A. Zussman (2020) "Identity and Bias: Insights from Driving Tests." *Economic Journal*, 130(625): 1-23.

Barron, K., R. Ditlmann, S. Gehrig, and S. Schweighofer-Kodritsch (2024) "Explicit and Implicit Belief-Based Gender Discrimination: A Hiring Experiment." *Management Science*.

Becker, G. (1957) *The economics of discrimination*. Chicago: University of Chicago Press.

Bertrand, M., and E. Duflo (2017) "Chapter 8 – Field Experiments on Discrimination." in *Handbook of Field Experiments*: 1: 309-393.

Bohren, J. A., A. Imas, and M. Rosenberg (2019) "The dynamics of discrimination: Theory and evidence." *American Economic Review*, 109(10): 3395-3436.

Bohren, J.A., P. Hull, and A. Imas (2022) "Systemic discrimination: Theory and measurement." NBER Working Paper no. 29820.

Bohren, J. A., K. Haggag, A. Imas, and D.G. Pope (2023) "Inaccurate Statistical Discrimination: An Identification Problem." *Review of Economics and Statistics*, 109(10): 1-45.

Bordalo, P., K. Coffman, N. Gennaioli, and A. Schleifer (2016) "Stereotypes." *Quarterly Journal of Economics*, 131(4): 1753-1794.

Bordalo, P., K. Coffman, N. Gennaioli, and A. Schleifer (2019) "Beliefs about Gender." *American Economic Review*, 109(3): 739-773.

Bertrand, M., and S. Mullainathan (2004) "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review*, 94(4): 991-1013.

Brown, M.B., and A.B. Forsythe (1974) "Robust tests for the equality of variances." *Journal of the American Statistical Association*, 69: 364-367.

Cappelen A.W., R. Falch, and B. Tungodden (2023) "Experimental evidence on the acceptance of males falling behind." NHH Department of Economics Discussion Paper No. 13/2023, Norwegian School of Economics, Bergen, Norway.

Cason, T. N., T. Sharma, and R. Vadovic (2020) "Correlated beliefs: Predicting outcomes in 2 × 2 games." *Games and Economic Behavior*, 122: 256-276.

Christensen, P., I. Sarmiento-Barbieri, and C. Timmins (2022) "Housing Discrimination and the Toxics Exposure Gap in the United States: Evidence from the Rental Market." *Review of Economics and Statistics*, 104(4): 807-818.

Clermont K.M., and S.J. Schwab (2009) "Employment discrimination plaintiffs in federal court: From bad to worse?" *Harvard Law & Policy Review,* 3(1): 103-32.

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

Coffman, K.B., C.L. Exley, and M. Niederle (2021) "The Role of Beliefs in Driving Gender Discrimination." *Management Science*, 67(6): 3551-3569.

Cunningham, T., and J. de Quidt (2024) "Implicit preferences inferred from choice." Working Paper.

De Paola, M., V. Scoppa (2015) "Gender discrimination and evaluators' gender: Evidence from Italian academia." *Economica*, 82(325): 162-188.

Dhar, D., T. Jain, and S. Jayachandran (2022) "Reshaping Adolescents' Gender Attitudes: Evidence from a School-Based Experiment in India." *American Economic Review*, 112(3): 899-927.

Dreber, A., E. von Essen, and E. Ranehill (2014) "Gender and competition in adolescence: task matters." *Experimental Economics*, 17: 154-172.

Ewens, M., B. Tomlin, and L.C. Wang (2014) "Statistical Discrimination or Prejudice? A Large Sample Field Experiment." *Review of Economics and Statistics*, 96(1): 119-134.

Foschi, M., L. Lai, and K. Sigerson (1994) "Gender and double standard in the assessment of job applicants." *Social Psychology Quarterly*, 57(4): 326-339.

Gaddis, S.M. (Ed.) (2018) *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*. Cham, Switzerland: Springer International Publishing.

Gaertner, S. L., and J. F. Dovidio (2000). *Reducing intergroup bias: The common ingroup identity model*. New York: Psychology Press.

Gawronski, B. (2019) "Six Lessons for a Cogent Science of Implicit Bias and Its Criticism." *Perspectives on Psychological Science*, 14(4): 574-595.

Hebl, M., S.K. Cheng, and L.C. Ng (2020) "Modern discrimination in organizations." *Annual Review of Organizational Psychology and Organizational Behavior*, 7: 257-282.

Hedegaard, M.S., and J. Tyran (2018) “The Price of Prejudice.” *American Economics Journal: Applied Economics,* 10(1): 40-63.

Heilman, M.E., and M.C. Haynes (2005): “No credit where credit is due: attributional rationalization of women’s success in male-female teams.” *Journal of Applied Psychology*, 90(5): 905.

Islam, A., D. Pakrashi, L.C. Wang, and Y. Zenou (2024) “Determining the Extent of Taste-Based and Accurate Statistical Discrimination: Evidence from a Field Experiment in India.” *Working paper.*

Levati M.V., A. Nicholas, and B. Rai (2014) “Testing the single-peakedness of other-regarding preferences.” *European Economic Review*, 67: 197-209.

List, J. A. (2004) “The Nature and Extent of Discrimination in the Marketplace: Evidence from the Field.” *Quarterly Journal of Economics*, 119(1): 49-89.

Levene, H. (1960) “Robust tests for equality of variances.” In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, and H. B. Mann (ed.), 278-292. Menlo Park, CA: Stanford University Press.

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

Moss-Racusin, C.A., J.F. Dovidio, V.L. Brescoll, M.J. Graham, and J. Handelsman (2012) “Science faculty’s subtle gender biases favor male students.” *Proceedings of the National Academy of Sciences*, 109(41): 16474–16479.

Neumark, D. (2012) “Detecting Discrimination in Audit and Correspondence Studies.” *Journal of Human Resources*, 47(4): 1128-1157.

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

Norton, M.I., J.A. Vandello, and J.M. Darley (2004) “Casuistry and Social Category Bias.” *Journal of Personality and Social Psychology*, 87(6): 817-831.

Pew Research Center (2021) “STEM Jobs See Uneven Progress in Increasing Gender, Racial and Ethnic Diversity.” https://www.pewresearch.org/social-trends/2021/04/01/stem-jobs-see-uneven-progress-in-increasing-gender-racial-and-ethnic-diversity/, accessed 3 June 2024.

Phelps, E.S. (1972) “The Statistical Theory of Racism and Sexism.” *American Economic Review*, 62(4): 659-661.

Pikulina, E.S., and D. Ferreira (2023) “Subtle Discrimination.” *Working paper.*

Riach, P.A., and J. Rich (2002) “Field Experiments of Discrimination in the Market Place.” *Economic Journal*, 112(483): F480-F518.

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

Reynolds, W.M. (1982) “Development of reliable and valid short forms of the Marlowe-Crowne Social Desirability Scale.” *Journal of Clinical Psychology*, 38(1): 119-125.

Schwartzstein, J. (2014) “Selective Attention and Learning.” *Journal of the European Economic Association*, 12(6): 1423-1452.

Uhlmann, E.L., and G.L. Cohen (2005) “Constructed Criteria: Redefining Merit to Justify Discrimination.” *Psychological Science*, 16(6): 474-480.

Uhlmann, E. L., and G.L. Cohen (2007) “"I think it, therefore it's true": Effects of self-perceived objectivity on hiring discrimination.” *Organizational Behavior and Human Decision Processes*, 104(2): 207–223.

Appendix: Additional tables

Table A.1 – Summary statistics and balance tests

Participant characteristic	(1) ACS 2022	(2) Choice treatment	(3) Ordinal belief treatment	(4) Cardinal belief treatment	(5) *p*-value
Male	0.502	0.500	0.497	0.500	0.935
		(0.041)	(0.041)	(0.041)	
Age 43 years or over	0.528	0.340	0.342	0.349	0.972
		(0.039)	(0.039)	(0.039)	
College degree or higher	0.465	0.460	0.477	0.467	0.644
		(0.041)	(0.041)	(0.041)	
Fraction of experimental instruction questions answered correctly		0.932	0.938	0.952	0.313
		(0.012)	(0.012)	(0.010)	
Fraction of real-effort questions answered correctly		0.893	0.879	0.898	0.982
		(0.020)	(0.022)	(0.017)	
Fraction of verbal questions answered correctly		0.318	0.322	0.344	0.176
		(0.019)	(0.019)	(0.021)	
Fraction of math questions answered correctly		0.328	0.346	0.316	0.938
		(0.022)	(0.023)	(0.022)	
Belief about own fraction of real-effort questions answered correctly		0.957	0.932	0.954	0.449
		(0.012)	(0.016)	(0.013)	
Belief about own fraction of verbal questions answered correctly		0.393	0.403	0.383	0.368
		(0.022)	(0.021)	(0.022)	
Belief about own fraction of math questions answered correctly		0.435	0.435	0.408	0.677
		(0.024)	(0.023)	(0.022)	
Fraction of attention check questions answered correctly		0.753	0.826	0.757	0.998
		(0.019)	(0.016)	(0.019)	
Republican-leaning		0.353	0.389	0.336	0.480
		(0.039)	(0.040)	(0.038)	
Supervisor		0.633	0.550	0.651	0.488
		(0.039)	(0.041)	(0.039)	
Found survey difficult		0.280	0.282	0.316	0.519
		(0.037)	(0.037)	(0.038)	
Online study experience		0.933	0.953	0.980	0.055**
		(0.037)	(0.034)	(0.028)	
Interested in math		0.587	0.584	0.533	0.551
		(0.040)	(0.041)	(0.041)	

Social desirability score (total score out of 13)	6.160	6.382	6.257	0.372
	(0.269)	(0.289)	(0.279)	
Perception of Kardashian knowledge (between -1 and 1)	-0.635	-0.704	-0.638	0.719
	(0.036)	(0.036)	(0.036)	
Perception of verbal skills knowledge (between -1 and 1)	-0.248	-0.270	-0.232	0.748
	(0.037)	(0.034)	(0.033)	
Perception of cars knowledge (between -1 and 1)	0.649	0.657	0.679	0.158
	(0.026)	(0.031)	(0.022)	
Perception of cooking knowledge (between -1 and 1)	-0.335	-0.346	-0.257	0.423
	(0.038)	(0.035)	(0.029)	
Perception of math knowledge (between -1 and 1)	0.171	0.175	0.204	0.293
	(0.031)	(0.033)	(0.031)	
Perception of Kardashian confidence (between -1 and 1)	-0.630	-0.663	-0.626	0.736
	(0.035)	(0.034)	(0.036)	
Perception of verbal skills confidence (between -1 and 1)	-0.118	-0.148	-0.122	0.606
	(0.041)	(0.039)	(0.038)	
Perception of cars confidence (between -1 and 1)	0.757	0.777	0.750	0.417
	(0.022)	(0.027)	(0.021)	
Perception of cooking confidence (between -1 and 1)	-0.383	-0.366	-0.281	0.377
	(0.040)	(0.039)	(0.036)	
Perception of math confidence (between -1 and 1)	0.307	0.296	0.278	0.302
	(0.035)	(0.035)	(0.033)	
F-statistic				0.690
Obs.	150	149	152	

Notes: Participant characteristics are reported as means. American Community Survey (ACS) 2022 data are based on the U.S. population aged between 25 and 64 years. Column 5 reports the *p*-value from the F-test of joint significance, which tests whether the set of characteristics jointly explain treatment status. Standard errors of the means clustered at the participant level are reported in parentheses. *** *p*<0.01; ** *p*<0.05; * *p*<0.10

Table A.2 – Differences in performance scores between groups

	(1) Real effort	(2) Verbal	(3) Math
MC = FC	-0.037	-0.025	0.059
	(0.245)	(0.333)	(0.207)
MN = FN	0.065**	-0.002	0.030
	(0.024)	(0.848)	(0.382)
MC = MN	-0.079***	-0.021	0.116**
	(0.009)	(0.385)	(0.011)
MC = FN	-0.015	-0.025	0.146***
	(0.674)	(0.336)	(0.001)
FC = MN	-0.043	0.003	0.056
	(0.139)	(0.871)	(0.188)
FC = FN	0.022	0.001	0.086**
	(0.443)	(0.971)	(0.036)
Obs.	451	451	451

Notes: Pools data for all participants. We construct a measure of average performance in a task category for each individual by first calculating the average share of correct answers (between 0 and 1) over all questions in the category. Then, we take the population average of this average performance for each of the 4 groups: (1) male participants with a college degree (MC), (2) female participants with a college degree (FC), (3) male participants with no college degree (MN), and (4) female participants with no college degree (FN). We then take the difference between the first group and second group's average performance (for example, the first row reports the difference MC – FC). *p*-value is given for the null hypothesis of no difference in average performance between two groups using the Wilcoxon-Mann Whitney rank-sum test for two samples. *p*-value reported in parentheses. *** *p*<0.01; ** *p*<0.05; * *p*<0.10

Table A.3 – Performance scores and rankings of groups

	(1) Real effort	(2) Verbal	(3) Math
A. Performance score			
MC	0.854	0.309	0.414
FC	0.891	0.335	0.354
MN	0.934	0.331	0.298
FN	0.869	0.333	0.268
Obs.	451	451	451
B. Performance ranking			
MC	4	4	1
FC	2	1	2
MN	1	3	3
FN	3	2	4
Obs.	451	451	451

Notes: Pools data for all participants. For performance scores, we calculate for each individual the average share of questions answered correctly (between 0 and 1) over all questions in the category. Then, we take the population average of this average performance for each of the 4 groups: (1) male participants with a college degree (MC), (2) female participants with a college degree (FC), (3) male participants with no college degree (MN), and (4) female participants with no college degree (FN). Performance rankings are based on the average share of questions answered correctly by a group (where 1 is the highest rank and 4 is the lowest rank).

Table A.4 – Frequencies of MC and FC groups being ranked first in Ordinal belief treatment

	(1) Real effort	(2) Verbal	(3) Math
A. All evaluators			
MC non-unique	0.721	0.582	0.631
FC non-unique	0.752	0.614	0.626
MC unique	0.084	0.112	0.205
FC unique	0.104	0.213	0.102
Obs.	149	149	149
B. Male evaluators			
MC non-unique	0.750	0.541	0.615
FC non-unique	0.716	0.584	0.608
MC unique	0.115	0.152	0.247
FC unique	0.095	0.229	0.091
Obs.	74	74	74
C. Female evaluators			
MC non-unique	0.693	0.623	0.647
FC non-unique	0.787	0.643	0.643
MC unique	0.053	0.073	0.163
FC unique	0.113	0.197	0.113
Obs.	75	75	75

Notes: For each individual, we construct a measure of male participants with a college degree (MC) and female participants with a college degree (FC) being tied for first rank for each individual by the average fraction of times (between 0 and 1) that MC is non-uniquely ranked first, and the average fraction of times FC is non-uniquely ranked first, on a question over all questions in the category. Then, we take the population average of this individual average across all participants in the Ordinal belief treatment. We construct a measure of strict first ranking in a task category for each individual by calculating the average fraction of times (between 0 and 1) that MC and FC are uniquely ranked first over all questions in the category. Then, we take the population average of this individual average across all participants in the Ordinal belief treatment.

Robustness check: Relative ranking between the MC and FC groups

We check whether the tie-breaking patterns observed for absolute first ranking also translate into tie-breaking patterns for relative ranking. For each individual, we construct binary variables which capture beliefs of three possible relative rankings: (i) whether the MC and FC groups have equal rank (ii) whether the MC group is ranked strictly higher than the FC group (iii) whether the FC group is ranked strictly higher than the MC group. We construct the same variables for choices, except we omit category (i) because ties are not allowed in the choice treatment.

As shown in Appendix Tables A.5 to A.7, we find the results for relative ranking are similar to those for absolute first ranking, albeit a bit weaker. The main difference is that we do not observe ties between the MC and FC groups being significantly broken in favor of men in math and in favor of women in real-effort tasks at the aggregate level. However, we still observe this tie-breaking pattern among male and female employers respectively, and that ties are broken in favor of women in verbal tasks regardless of the employer's gender. This suggests that the bias in tie-breaking is relatively larger for ties at first rank than for ties at lower ranks.

Table A.5 – Relative ranking between MC and FC groups in the Choice treatment

	(1) Real effort	(2) Verbal	(3) Math
A. All employers			
MC higher rank than FC	0.443	0.330	0.583
FC higher rank than MC	0.557	0.670	0.417
MC higher rank - FC higher rank	-0.113*	-0.340***	0.167***
	(0.084)	(0.000)	(0.009)
Obs.	150	150	150
B. Male employers			
MC higher rank than FC	0.520	0.363	0.630
FC higher rank than MC	0.480	0.637	0.370
MC higher rank - FC higher rank	0.040	-0.273***	0.260***
	(0.668)	(0.004)	(0.006)
Obs.	75	75	75
C. Female employers			
MC higher rank than FC	0.367	0.297	0.537
FC higher rank than MC	0.633	0.703	0.463
MC higher rank - FC higher rank	-0.267***	-0.407***	0.073
	(0.004)	(0.000)	(0.395)
Obs.	75	75	75

Notes: We construct a measure of relative ranking in a task category for each individual by calculating the average fraction of times (between 0 and 1) that male participants with a college degree (MC) strictly ranked higher than female participants with a college degree (FC), and the average fraction of times FC is strictly ranked higher than MC. Then, we take the population average of this individual average across all participants in the Choice treatment, and take the difference between the MC and FC averages (choice gap). *p*-value is given for the null hypothesis of no difference in the average fraction of times MC is strictly ranked higher than FC and the average fraction of times FC is ranked higher than MC using the Wilcoxon signed-rank test on matched pairs. *p*-value reported in parentheses. *** p<0.01; ** p<0.05; * p<0.10

Table A.6 – Evaluators' relative ranking between MC and FC groups in the Ordinal belief treatment

	(1) Real effort	(2) Verbal	(3) Math
A. All evaluators			
MC and FC equal rank	0.695	0.565	0.601
MC higher rank than FC	0.131	0.151	0.253
FC higher rank than MC	0.174	0.284	0.146
MC higher rank - FC higher rank	-0.044	-0.133***	0.107***
	(0.268)	(0.000)	(0.004)
Obs.	149	149	149
B. Male evaluators			
MC and FC equal rank	0.696	0.534	0.588
MC higher rank than FC	0.176	0.172	0.291
FC higher rank than MC	0.128	0.294	0.122
MC higher rank - FC higher rank	0.047	-0.122**	0.169***
	(0.153)	(0.017)	(0.007)
Obs.	74	74	74
C. Female evaluators			
MC and FC equal rank	0.693	0.597	0.613
MC higher rank than FC	0.087	0.130	0.217
FC higher rank than MC	0.220	0.273	0.170
MC higher rank - FC higher rank	-0.133***	-0.143***	0.047
	(0.006)	(0.005)	(0.206)
Obs.	75	75	75

Notes: We construct a measure of relative ranking in a task category for each individual by calculating the average fraction of times (between 0 and 1) that male participants with a college degree (MC) and female participants with a college degree (FC) have the same rank, the average fraction of times MC is strictly ranked higher than FC, and the average fraction of times FC is strictly ranked higher than MC. Then, we take the population average of this individual average across all participants in the Ordinal belief treatment, and take the difference between the MC and FC averages. *p*-value is given for the null hypothesis of no difference in the average fraction of times MC is ranked higher than FC and the average fraction of times FC is ranked higher than MC using the Wilcoxon signed-rank test on matched pairs. *p*-value reported in parentheses. *** p<0.01; ** p<0.05; * p<0.10

Table A.7 – Excess hiring gap between MC and FC groups across task categories based on relative rankings

	(1) Real effort	(2) Verbal	(3) Math
A. All participants			
Excess hiring gap	-0.069	-0.207***	0.059
	(0.212)	(0.000)	(0.102)
Obs. (Choice treatment)	150	150	150
Obs. (Cardinal belief)	152	152	152
B. Male participants			
Excess hiring gap	-0.007	-0.152**	0.091*
	(0.953)	(0.029)	(0.079)
Obs. (Choice treatment)	75	75	75
Obs. (Cardinal belief)	76	76	76
C. Female participants			
Excess hiring gap	-0.133*	-0.263***	0.027
	(0.073)	(0.002)	(0.686)
Obs. (Choice treatment)	75	75	75
Obs. (Cardinal belief)	76	76	76

Notes: p-value reported in parentheses. *** p<0.01; ** p<0.05; * p<0.10

Robustness check: Beliefs about gender gaps in performance by Evaluators in the Cardinal belief treatment

For our main results, we compare the choices of participants with the beliefs of participants in the ordinal belief treatment. One possible concern is that the observed believed gender gaps in performance and therefore results may depend on the method of belief elicitation. For example, gender comparisons may be more salient in ordinal rankings than in cardinal numbers. We can explore the extent of this issue by using the alternative beliefs of participants in the cardinal belief treatment. Recall that the cardinal belief treatment asked participants to guess the share of participants in a group (from 0 to 100) who answered a question correctly. To compare the cardinal beliefs with ordinal beliefs, we convert the cardinal beliefs to their implicit ordinal ranks.

We first ask how belief gaps vary by the belief elicitation method. Using the rank sum test, we find no significant differences in belief gaps between the cardinal belief and ordinal belief treatment across all domains (Appendix Table A.3). Therefore, it is unlikely participants are more likely to report belief gaps in their ordinal beliefs than in their cardinal beliefs. We then check whether our main results also hold for cardinal beliefs. In Appendix Tables A.8 – A.14, we show that our main results are robust if we replace ordinal belief gaps with cardinal belief gaps.

Although cardinal and ordinal belief gaps are observationally similar when cardinal beliefs are converted to ordinal ranks, measured belief gaps are much weaker in cardinal beliefs than ordinal beliefs if we use beliefs in its direct cardinal form. Appendix Table A.12 shows that direct cardinal beliefs only detect significant believed gaps in verbal tasks but not in math tasks, whereas ordinal beliefs detect significant believed gaps in both verbal and math tasks.

Therefore, it may be important to consider both cardinal and ordinal measures of beliefs to determine whether there are believed performance differences between genders.

Table A.8 – Difference in the frequencies of the MC and FC groups being uniquely ranked first in belief-rankings of evaluators in the Cardinal belief treatment

	(1) Real effort	(2) Verbal	(3) Math
A. All evaluators			
Belief gap	-0.013	-0.115**	0.082*
	(0.759)	(0.018)	(0.069)
Obs.	152	152	152
B. Male evaluators			
Belief gap	0.013	-0.158**	0.033
	(0.672)	(0.025)	(0.613)
Obs.	76	76	76
C. Female evaluators			
Belief gap	-0.039	-0.072	0.132**
	(0.295)	(0.303)	(0.027)
Obs.	76	76	76

Notes: We first convert cardinal beliefs of performance to their implicit ordinal rankings. We construct a measure of strict first ranking in a task category for each individual by calculating the average fraction of times (between 0 and 1) that male participants with a college degree (MC) and female participants with a college degree (FC) are uniquely ranked first over all questions in the category. Then, we take the population average of this individual average across all participants in the Cardinal belief treatment and take the difference between the MC and FC averages for the strict rankings (MC – FC). *p*-value is given for the null hypothesis of no difference in the average fraction of times ranked first between MC and FC using the Wilcoxon signed-rank test on matched pairs. *p*-value reported in parentheses. *** $p<0.01$; ** $p<0.05$; * $p<0.10$

Table A.9 – Difference in the frequencies of the MC and FC groups being both ranked first by evaluators in the Ordinal and Cardinal belief treatments

	(1) Ties in first rank in Ordinal belief treatment (MC – FC)	(2) Ties in first rank in Cardinal belief treatment (MC – FC)	(3) *p*-value
A. All evaluators			
Real effort	0.688	0.655	0.422
Verbal	0.555	0.389	0.000
Math	0.594	0.399	0.000
Obs.	149	152	301
B. Male evaluators			
Real effort	0.696	0.651	0.340
Verbal	0.527	0.356	0.009
Math	0.584	0.332	0.000
Obs.	74	76	150
C. Female evaluators			
Real effort	0.680	0.658	0.844
Verbal	0.583	0.424	0.012
Math	0.603	0.467	0.023
Obs.	75	76	151

Notes: We first convert cardinal beliefs of performance to their implicit ordinal rankings. We first calculate the total number of times male participants with a college degree (MC) and female participants with a college degree (FC) are both ranked first by an individual participant in the Ordinal belief treatment and Cardinal belief treatment across real effort (out of 2), verbal (out of 4), and math (out of 4). We then calculate the average number of times (i.e. fraction of questions in each category) MC and FC are ranked 1st at the individual level (between 0 and 1). Finally, we take the individual average across all participants in the Ordinal belief treatment and Cardinal belief treatment to give Columns 1 and 2 respectively. Column 3 gives the *p*-value for the null hypothesis of no difference in the distribution in the average gender gap in the fraction of times MC and FC are both ranked first between the Ordinal belief treatment and Cardinal belief treatment using the Wilcoxon-Mann-Whitney rank-sum test for two samples. *p*-value reported in parentheses. *** p<0.01; ** p<0.05; * p<0.10

Table A.10 – Difference in the frequencies of the MC and FC groups being ranked first in strict rankings of evaluators in the Ordinal and Cardinal belief treatments

	(1) Uniquely ranked first in Ordinal belief treatment (MC – FC)	(2) Uniquely ranked first in Cardinal belief treatment (MC – FC)	(3) *p*-value
A. All evaluators			
Real effort	-0.020	-0.013	0.853
Verbal	-0.101	-0.115	0.910
Math	0.102	0.082	0.802
Obs.	149	152	301
B. Male evaluators			
Real effort	0.020	0.013	0.969
Verbal	-0.078	-0.158	0.363
Math	0.155	0.033	0.274
Obs.	74	76	150
C. Female evaluators			
Real effort	-0.060	-0.039	0.784
Verbal	-0.123	-0.072	0.414
Math	0.050	0.132	0.397
Obs.	75	76	151

Notes: We first convert cardinal beliefs of performance to their implicit ordinal rankings. We first calculate the total number of times male participants with a college degree (MC) and female participants with a college degree (FC) are strictly ranked first by an individual participant in the Ordinal belief treatment and Cardinal belief treatment across real effort (out of 2), verbal (out of 4), and math (out of 4). We then calculate the average number of times (i.e. fraction of questions in each category) MC and FC are ranked 1st at the individual level (between 0 and 1). Finally, we take the individual average across all participants in the Ordinal belief treatment and Cardinal belief treatment to give Columns 1 and 2 respectively. Column 3 gives the *p*-value for the null hypothesis of no difference in the distribution in the average gender gap in the fraction of times MC and FC are strictly ranked first between the Ordinal belief treatment and Cardinal belief treatment using the Wilcoxon-Mann-Whitney rank-sum test for two samples. *p*-value reported in parentheses. *** p<0.01; ** p<0.05; * p<0.10

Table A.11 – Excess hiring gap between MC and FC groups across task categories based on first rank

	(1) Real effort	(2) Verbal	(3) Math
A. All participants			
Excess hiring gap	-0.107*	-0.208***	0.106*
	(0.089)	(0.001)	(0.060)
Obs. (Choice treatment)	150	150	150
Obs. (Cardinal belief)	152	152	152
B. Male participants			
Excess hiring gap	0.027	-0.092	0.244**
	(0.704)	(0.193)	(0.016)
Obs. (Choice treatment)	75	75	75
Obs. (Cardinal belief)	76	76	76
C. Female participants			
Excess hiring gap	-0.241***	-0.324***	-0.032
	(0.004)	(0.000)	(0.962)
Obs. (Choice treatment)	75	75	75
Obs. (Cardinal belief)	76	76	76

Notes: p-value reported in parentheses. *** p<0.01; ** p<0.05; * p<0.10

Table A.12 – Evaluators' beliefs about performance scores of the MC and FC groups in the Cardinal belief treatment

	(1) Performance score (MC)	(2) Performance score (FC)	(3) Performance gap (MC – FC)	(4) *p*-value
A. All evaluators				
Real effort	0.904	0.900	0.004	0.647
Verbal	0.675	0.691	-0.016***	0.002
Math	0.745	0.732	0.014	0.127
Obs.	152	152	152	152
B. Male evaluators				
Real effort	0.889	0.884	0.005	0.352
Verbal	0.635	0.650	-0.015**	0.012
Math	0.715	0.709	0.007	0.559
Obs.	76	76	76	76
C. Female evaluators				
Real effort	0.919	0.916	0.003	0.720
Verbal	0.716	0.732	-0.016**	0.049
Math	0.775	0.755	0.020	0.131
Obs.	76	76	76	76

Notes: For each individual, we take the average of the guessed percentage of correct answers (between 0 and 100) for males with college degrees (MC) and females with college degrees (FC) across real effort, verbal and math questions, and then divide by 100 to get a fraction which lies between 0 and 1. Then, we take the population average of this individual average across all participants in the Cardinal belief treatment, and take the difference between the MC and FC averages (MC – FC). *p*-value is given for the null hypothesis of no difference in the average guessed fraction of correct answers between MC and using the Wilcoxon signed-rank test on matched pairs. *p*-value reported in parentheses. *** p<0.01; ** p<0.05; * p<0.10

Table A.13 – Evaluators' relative ranking between the MC and FC groups in the Cardinal belief treatment

	(1) Real effort	(2) Verbal	(3) Math
A. All evaluators			
MC and FC equal rank	0.602	0.321	0.352
MC higher rank than FC	0.191	0.406	0.291
FC higher rank than MC	0.207	0.273	0.357
MC higher rank - FC higher rank	-0.016	0.133***	-0.066
	(0.518)	(0.008)	(0.176)
Obs.	152	152	152
B. Male evaluators			
MC and FC equal rank	0.599	0.303	0.289
MC higher rank than FC	0.171	0.428	0.355
FC higher rank than MC	0.230	0.269	0.355
MC higher rank - FC higher rank	-0.059	0.158**	0.000
	(0.208)	(0.029)	0.967
Obs.	76	76	76
C. Female evaluators			
MC and FC equal rank	0.605	0.339	0.414
MC higher rank than FC	0.211	0.385	0.227
FC higher rank than MC	0.184	0.276	0.359
MC higher rank - FC higher rank	0.026	0.109	-0.132**
	(0.674)	(0.116)	(0.039)
Obs.	76	76	76

Notes: We first convert cardinal beliefs of performance to their implicit ordinal rankings. We construct a measure of relative ranking in a task category for each individual by calculating the average fraction of times (between 0 and 1) that male participants with a college degree (MC) and female participants with a college degree (FC) have the same rank, average fraction of times MC is strictly ranked higher than FC, and average fraction of times FC is strictly ranked higher than MC. Then, we take the population average of this individual average across all participants in the Cardinal belief treatment, and take the difference between the MC and FC averages (MC – FC). This gives us the "belief gap." *p*-value is given for the null hypothesis of no difference in the average fraction of times MC is ranked higher than FC and the average fraction of times FC is ranked higher than MC using the Wilcoxon signed-rank test on matched pairs. *p*-value reported in parentheses. *** $p<0.01$; ** $p<0.05$; * $p<0.10$

Table A.14 – Excess hiring gap between MC and FC groups across task categories based on relative rankings

	(1) Real effort	(2) Verbal	(3) Math
A. All participants			
Excess hiring gap	-0.097	-0.473***	0.232**
	(0.144)	(0.000)	(0.001)
Obs. (Choice treatment)	150	150	150
Obs. (Cardinal belief)	152	152	152
B. Male participants			
Excess hiring gap	0.099	-0.431***	0.260**
	(0.385)	(0.000)	(0.011)
Obs. (Choice treatment)	75	75	75
Obs. (Cardinal belief)	76	76	76
C. Female participants			
Excess hiring gap	-0.293	-0.515***	0.205*
	(0.003)	(0.000)	(0.068)
Obs. (Choice treatment)	75	75	75
Obs. (Cardinal belief)	76	76	76

Notes: p-value reported in parentheses. *** *p*<0.01; ** *p*<0.05; * *p*<0.10

A.15 – Employers' relative ranking of the MN versus FN groups in the Choice treatment

	(1) Real effort	(2) Verbal	(3) Math
A. All evaluators			
MN higher rank than FN	0.387	0.332	0.532
FN higher rank than MN	0.613	0.668	0.468
MN higher rank – FN higher rank	-0.227***	-0.337***	0.063
	(0.001)	(0.000)	(0.335)
Obs.	150	150	150
B. Male evaluators			
MN higher rank than FN	0.433	0.360	0.547
FN higher rank than MN	0.567	0.640	0.453
MN higher rank - FN higher rank	-0.133	-0.280***	0.093
	(0.140)	(0.002)	(0.333)
Obs.	75	75	75
C. Female evaluators			
MN higher rank than FN	0.340	0.303	0.517
FN higher rank than MN	0.660	0.697	0.483
MN higher rank - FN higher rank	-0.320***	-0.393***	0.033
	(0.001)	(0.000)	(0.709)
Obs.	75	75	75

Notes: We construct a measure of relative ranking in a task category for each individual by calculating the average fraction of times (between 0 and 1) that males with no college degree (MN) is strictly ranked higher than females with no college degree (FN), and the average fraction of times FN is strictly ranked higher than MN. Then, we take the population average of this individual average across all participants in the Choice treatment, and take the difference between the MN and FN averages (MN – FN). This gives us the "choice gap". *p*-value is given for the null hypothesis of no difference in the average fraction of times MN is ranked higher than FN and the average fraction of times FN is ranked higher than MN using the Wilcoxon signed-rank test on matched pairs. *p*-value reported in parentheses. *** p<0.01; ** p<0.05; * p<0.10

Table A.16 – Evaluators' relative ranking of MN versus FN groups in the Ordinal belief treatment

	(1) Real effort	(2) Verbal	(3) Math
A. All evaluators			
MN and FN equal rank	0.654	0.592	0.594
MN higher rank than FN	0.134	0.129	0.221
FN higher rank than MN	0.211	0.279	0.185
MN higher rank - FN higher rank	-0.077**	-0.149***	0.037
	(0.047)	(0.000)	(0.267)
Obs.	149	149	149
B. Male evaluators			
MN and FN equal rank	0.649	0.608	0.571
MN higher rank than FN	0.182	0.145	0.267
FN higher rank than MN	0.169	0.247	0.162
MN higher rank - FN higher rank	0.014	-0.101*	0.105*
	(0.638)	(0.059)	(0.067)
Obs.	74	74	74
C. Female evaluators			
MN and FN equal rank	0.660	0.577	0.617
MN higher rank than FN	0.087	0.113	0.177
FN higher rank than MN	0.253	0.310	0.207
MN higher rank - FN higher rank	-0.167***	-0.197***	-0.030
	(0.001)	(0.000)	(0.703)
Obs.	75	75	75

Notes: We construct a measure of relative ranking in a task category for each individual by calculating the average fraction of times (between 0 and 1) that males with no college degree (MN) and females with no college degree (FN) have equal rank, the average fraction of times MN is strictly ranked higher than FN, and the average fraction of times FN is strictly ranked higher than MN. Then, we take the population average of this individual average across all participants in the Ordinal belief treatment, and take the difference between the MN and FN averages (MN – FN). This gives us the "belief gap." *p*-value is given for the null hypothesis of no difference in the average fraction of times MN is ranked higher than FN, and the average fraction of times FN is ranked higher than MN using the Wilcoxon signed-rank test on matched pairs. *p*-value reported in parentheses. *** p<0.01; ** p<0.05; * p<0.10

Table A.17 – Excess hiring gap between MN and FN groups across task categories based on relative rankings

	(1) Real effort	(2) Verbal	(3) Math
A. All participants			
Excess hiring gap	-0.149**	-0.187***	0.026
	(0.012)	(0.001)	(0.468)
Obs. (Choice treatment)	150	150	150
Obs. (Ordinal belief treatment)	149	149	149
B. Male participants			
Excess hiring gap	-0.147	-0.179**	-0.011
	(0.129)	(0.036)	(0.738)
Obs. (Choice treatment)	75	75	75
Obs. (Ordinal belief treatment)	74	74	74
C. Female participants			
Excess hiring gap	-0.153**	-0.197***	0.063
	(0.032)	(0.008)	(0.471)
Obs. (Choice treatment)	75	75	75
Obs. (Ordinal belief treatment)	75	75	75

Notes: *p*-value reported in parentheses. *** p<0.01; ** p<0.05; * p<0.10

Table A.18 – Evaluators' relative ranking of MN versus FN groups in Cardinal belief treatment

	(1) Real effort	(2) Verbal	(3) Math
A. All evaluators			
MN and FN equal rank	0.572	0.298	0.303
MN higher rank than FN	0.237	0.454	0.368
FN higher rank than MN	0.191	0.258	0.329
MN higher rank - FN higher rank	0.046	0.206***	0.039
	(0.374)	(0.000)	(0.454)
Obs.	152	152	152
B. Male evaluators			
MN and FN equal rank	0.566	0.316	0.269
MN higher rank than FN	0.250	0.474	0.408
FN higher rank than MN	0.184	0.211	0.322
MN higher rank - FN higher rank	0.066	0.263***	0.086
	(0.391)	(0.000)	(0.228)
Obs.	76	76	76
C. Female evaluators			
MN and FN equal rank	0.579	0.279	0.336
MN higher rank than FN	0.224	0.434	0.329
FN higher rank than MN	0.197	0.286	0.336
MN higher rank - FN higher rank	0.026	0.148**	-0.007
	(0.692)	(0.039)	0.887
Obs.	76	76	76

Notes: We first convert cardinal beliefs of performance to their implicit ordinal rankings. We construct a measure of relative ranking in a task category for each individual by calculating the average fraction of times (between 0 and 1) that males with no college degree (MN) and females with no college degree (FN) have equal rank, the average fraction of times MN is strictly ranked higher than FN, and the average fraction of times FN is strictly ranked higher than MN. Then, we take the population average of this individual average across all participants in the Ordinal belief treatment, and take the difference between the MN and FN averages (MN – FN). This gives us the "belief gap." *p*-value is given for the null hypothesis of no difference in the average fraction of times MN is ranked higher than FN, and the average fraction of times FN is ranked higher than MN using the Wilcoxon signed-rank test on matched pairs. *p*-value reported in parentheses. *** *p*<0.01; ** *p*<0.05; * *p*<0.10

Table A.19 – Excess hiring gap between MN and FN groups across task categories based on relative rankings

	(1) Real effort	(2) Verbal	(3) Math
A. All participants			
Excess hiring gap	-0.272***	-0.542***	0.024
	(0.000)	(0.000)	(0.601)
Obs. (Choice treatment)	150	150	150
Obs. (Cardinal belief treatment)	152	152	152
B. Male participants			
Excess hiring gap	-0.199*	-0.543***	0.008
	(0.068)	(0.000)	(0.728)
Obs. (Choice treatment)	75	75	75
Obs. (Cardinal belief treatment	76	76	76
C. Female participants			
Excess hiring gap	-0.346***	-0.541***	0.039
	(0.001)	(0.000)	(0.672)
Obs. (Choice treatment)	75	75	75
Obs. (Cardinal belief treatment)	76	76	76

Notes: p-value reported in parentheses. *** *p*<0.01; ** *p*<0.05; * *p*<0.10

Table A.20 – Evaluators' relative ranking over MC and FC groups by Social Desirability Scores

	(1) All evaluators	(2) Male evaluators	(3) Female evaluators
A. Difference in proportions between low and high SD evaluators in Real effort			
MC = FC	-0.042	0.084	-0.169*
	(0.527)	(0.368)	(0.071)
MC > FC	-0.039	-0.091	0.019
	(0.451)	(0.213)	(0.602)
MC < FC	0.082	0.007	0.149*
	(0.231)	(0.787)	(0.090)
Obs.	149	74	75
B. Difference in proportions between low and high SD evaluators in Verbal			
MC = FC	0.068	0.188**	-0.057
	(0.334)	(0.045)	(0.469)
MC > FC	-0.036	-0.057	-0.011
	(0.602)	(0.428)	(0.881)
MC < FC	-0.032	-0.131*	0.069
	(0.403)	(0.086)	(0.579)
Obs.	149	74	75
C. Difference in proportions between low and high SD evaluators in Math			
MC = FC	0.012	0.049	-0.028
	(0.853)	(0.631)	(0.776)
MC > FC	-0.014	-0.029	0.008
	(0.995)	(0.858)	(0.757)
MC < FC	0.002	-0.020	0.019
	(0.622)	(0.555)	(0.302)
Obs.	149	74	75

Notes: The social desirability (SD) score is a measure of a participant's propensity to give socially desirability answers. High SD refers to having an above-median score among participants. The table reports the differences in the proportion of participants who gave equal ranking to males with college degrees (MC) and females with college degree (FC), differences in the proportion of participants who gave MC a strictly higher ranking than FC, and differences in the proportion of participants in the Ordinal belief treatment who gave FC a strictly higher ranking than MC between participants with low and high SD scores. *p*-value reported in parentheses. *p*-value is given for the null hypothesis of no average difference in the proportions of relative rankings between participants with low and high SD scores using the Wilcoxon-Mann Whitney rank-sum test for two samples. *** p<0.01; ** p<0.05; * p<0.10

Table A.21 – Evaluators' relative ranking over MN and FN groups by Social Desirability Scores

	(1) All evaluators	(2) Male evaluators	(3) Female evaluators
A. Difference in proportions between low and high SD evaluators in Real effort			
MN = FN	-0.049	0.019	-0.120
	(0.471)	(0.791)	(0.189)
MN > FN	-0.019	-0.023	-0.008
	(0.738)	(0.707)	(0.907)
MN < FN	0.069	0.004	0.128
	(0.187)	(0.856)	(0.141)
Obs.	149	74	75
B. Difference in proportions between low and high SD evaluators in Verbal			
MN = FN	0.009	0.021	0.002
	(0.899)	(0.813)	(0.982)
MN > FN	-0.036	-0.069	-0.001
	(0.629)	(0.446)	(0.922)
MN < FN	0.026	0.048	-0.001
	(0.711)	(0.705)	(0.964)
Obs.	149	74	75
C. Difference in proportions between low and high SD evaluators in Math			
MN = FN	-0.061	-0.078	-0.049
	(0.333)	0.403	(0.568)
MN > FN	0.028	0.019	0.045
	(0.305)	0.683	(0.281)
MN < FN	0.033	0.059	0.004
	(0.449)	(0.711)	(0.579)
Obs.	149	74	75

Notes: The social desirability (SD) score is a measure of a participant's propensity to give socially desirability answers. High SD refers to having an above-median score among participants. The table reports the differences in the proportion of participants who gave equal ranking to males with no college degrees (MN) and females with no college degree (FN), differences in the proportion of participants who gave MN a strictly higher ranking than FN, and differences in the proportion of participants in the Ordinal belief treatment who gave FN a strictly higher ranking than MN between participants with low and high SD scores. *p*-value reported in parentheses. *p*-value is given for the null hypothesis of no average difference in the proportions of relative rankings between participants with low and high SD scores using the Wilcoxon-Mann Whitney rank-sum test for two samples. *** p<0.01; ** p<0.05; * p<0.10