The Role of Beliefs in Driving Gender Discrimination

Katherine B. Coffman,
Christine L. Exley,
Muriel Niederle *†

February 28, 2020

Abstract

While there is ample evidence of discrimination against women in the workplace, it can be difficult to understand what factors contribute to discriminatory behavior. We use an experiment to both document discrimination and unpack its sources. First, we show that, on average, employers prefer to hire male over female workers for a male-typed task even when the two workers have identical resumes. Second, and most critically, we use a control condition to identify that this discrimination is not specific to gender. Employers are simply less willing to hire a worker from a group that performs worse on average, even when this group is instead defined by a non-stereotypical characteristic. In this way, beliefs about average group differences are the key driver of discrimination against women in our setting. We also document some evidence for in-group preferences that contribute to the gender discrimination observed. Finally, our design allows us to understand and quantify the extent to which image concerns mitigate discriminatory behavior.

1 Introduction

Understanding the drivers of gender differences in labor market outcomes has been an important topic of study among labor economists, with research identifying sizable roles for occupational segregation, differences in human capital accumulation, demand for flexibility, and differences in preferences (Goldin, 2014; Card, Cardoso and Kline, 2016; Olivetti and Petrongolo, 2016).[1]

*Coffman: kcoffman@hbs.edu, Harvard Business School; Exley: clexley@hbs.edu, Harvard Business School; Niederle: niederle@stanford.edu, Stanford University, SIEPR and NBER.

†Thank you to John Beshears, Alison Wood Brooks, Lucas Coffman, Benjamin Edelman, Leslie John, Johanna Mollerstrom, Joshua Schwartzstein, Andrei Shleifer, and seminar audiences at Cornell University, the Rotman School of Management, and the 2016 ESA North American Meetings for their many useful comments.

[1]For reviews on gender differences, see Croson and Gneezy (2009), Bertrand (2011), Azmat and Petrongolo (2014) and Niederle (2016).

Discrimination also contributes to these gender gaps in earnings and advancement. A large body of empirical work has documented the existence of gender discrimination in labor markets, often through clever field experiments or with quasi-experimental data (for reviews, see Riach and Rich (2002) and Blau and Kahn (2017)). Discrimination against women has been found in bargaining contexts (Ayres and Siegelman, 1995; Castillo et al., 2013; Bowles, Babcock and Lai, 2007), in hiring, employment, and referral contexts (Neumark, Bank and Nort, 1996; Goldin and Rouse, 2000; Black and Strahan, 2001; Bertrand and Mullainathan, 2004; Moss-Racusin et al., 2012; Reuben, Sapienza and Zingales, 2014; Baert, De Pauw and Deschacht, 2016; Bohnet, van Geen and Bazerman, 2016; Sarsons, 2017*a*), and in academic contexts (Milkman, Akinola and Chugh, 2012, 2015; Sarsons, 2017*b*).[2]

Despite this large literature, important open questions about gender discrimination remain. One challenge is identifying the primary driver of observed discrimination. Our study attempts to answer this key question. Traditionally, economists have viewed discrimination through two distinct lenses: taste-based or statistical. The taste-based argument posits that discrimination is rooted in preferences, driven by animus or prejudice (Becker, 1957). Statistical discrimination, on the other hand, is rooted in rational beliefs about average gender differences in abilities or skills (Phelps, 1972; Arrow, 1973).[3] Recently, researchers have proposed more general forms of belief-based discrimination, with a particular focus on discrimination based upon inaccurate beliefs. Bordalo et al. (2016) document a role for representativeness-based stereotypes, where beliefs about groups are not accurate but instead are biased by differences in low probability but highly representative tails. These inaccurate beliefs can then give rise to self-stereotyping and discriminatory behavior (Coffman, 2014; Bordalo et al., 2019). In this spirit, Bohren, Imas and Rosenberg (2017) document discrimination against women on an online math platform and show that it is most consistent with discrimination based upon biased beliefs.

However, cleanly disentangling these different drivers of discrimination — particularly accurate beliefs, biased beliefs, and tastes — is challenging. Past literature has mainly followed three approaches. One approach is to test for a role for beliefs by increasing the amount of information available on individual workers. The argument is that if more informative performance information reduces discrimination, beliefs must play an important role. However, to the extent that discrimination remains, it is unclear whether it reflects taste-based considerations or residual belief-based motivations. A second approach is to simply remove information that identifies the (gender, ethnic, socio-economic, etc.) group to which an individual belongs, for

[2]Many scholars have identified discrimination against other groups as well, such as the beauty premium (Mobius and Rosenblat, 2006), discrimination based on socioeconomic status (Rao, 2018), and racial or ethnic discrimination (for examples, see Fershtman and Gneezy (2001), List (2004), Charles and Guryan (2008), Castillo and Petrie (2010), Gneezy (2012), Doleac and Stein (2013), Lei and Babcock (2013), and Bartoš et al. (2016), and, for a review, see Charles and Guryan (2011)). For a review, see also Bertrand and Duflo (2017).

[3]See Altonji and Blank (1999) and Guryan and Charles (2013) for review and discussion of literature on taste-based versus statistical discrimination.

instance by taking a name off of a resume, ensuring that any discrimination does not reflect taste-based motivations.[4] However, in contexts where there are group differences in performance on average, removing information that identifies the group to which an individual belongs not only rules out taste-based motivations, but also inhibits belief-based discrimination. A final approach, often used in field data, is to ask whether the observed discrimination is consistent with profit-maximization, suggesting statistical motivations. However, this has typically required assuming that the discrimination is driven by accurate, rather than inaccurate, beliefs.

Ideally, what one would do to separate any role of beliefs from tastes would be to recreate the exact hiring environment where we see discrimination against women, divorced from any gender-specific considerations but not ability-specific considerations. This is our goal. Of course, this goal would be nearly impossible to achieve in naturally occurring field data. And so, we design a controlled experiment. We employ a control condition that allows us to benchmark the level of discrimination we observe against women in a male-typed hiring environment to discrimination against workers of unknown gender known to be drawn from the same ability distribution as the women. We do this in a setting where we can provide accurate information both on individual workers and group differences and where we can measure employer beliefs. It is in this way that we attempt to separate discrimination based upon tastes and discrimination based upon beliefs (independent of whether those beliefs are accurate or not).

In a preliminary study, we collect performance information from participants on easy and hard math and sports quizzes. We use these participants as available "workers" for hire in our main hiring experiment. In this main hiring experiment, we ask "employers" to make incentivized choices over available workers. Employers receive information about the easy quiz performances of available workers. When they choose to hire a worker, they are paid based upon the hard round performance of that worker, which is unknown to them at the time of the decision.

The hiring experiment involves two treatments that vary only in their labeling of the available workers: a Gender treatment and a Birth Month treatment. In the Gender treatment, participants make hiring decisions between female and male workers, labeled as such, while in the Birth Month treatment, participants make decisions over workers born in an even month or workers born in an odd month. Importantly, the actual set of available workers is held constant across the two treatments; we simply vary how the workers are labeled. To implement this, we use two subsamples to generate our pool of available workers: women born in even months and men born in odd months. Thus, while the labels used to describe the available workers vary by treatment, the performances of the available workers do not. In both treatments, participants

[4] A distinct but related approach is to simulate a context where taste-based considerations are absent, for instance by studying anonymous workers with randomly-assigned productivities, allowing for clean documentation of statistical discrimination (for instance, Dickinson and Oaxaca (2009) and Dickinson and Oaxaca (2014)). Of course, absent a comparable context where taste-based considerations *are* relevant, this approach does not allow for one to back out how much of the observed discrimination is attributable to beliefs versus tastes.

receive accurate, detailed information about the distribution of performances across the groups prior to making their hiring decisions, generating similar ex-post beliefs about the average ability gap (male/female gap and odd/even gap) across the two treatments. In this way, we abstract away from the belief formation process.[5] Importantly, what we care most about is not whether beliefs are accurate or not (although our data can speak to this), but instead on how a given set of beliefs drives behavior in each treatment.

Our design creates a very similar set of beliefs about average group differences across our two treatments (the Gender treatment and the Birth Month treatment). Thus, we can ask whether women are more or less likely to be hired when they are labeled as women, in the Gender treatment, than when they are not, in the Birth Month treatment. Because beliefs of not only individual performances but also average group differences are the same across the two treatments, differential discrimination across the two treatments can be attributable to gender-specific tastes, rather than beliefs.

Our new methodology has a number of advantages. First, we elicit employer beliefs of gender differences, allowing us to directly link beliefs to behavior. Second, our Birth Month treatment better isolates the gender-specific taste component of discrimination by holding beliefs, both about individuals and groups, constant. It does so by removing information about worker gender while simultaneously maintaining identical information about individual and group-level performances. In this way, it provides a useful benchmark to which we can compare the extent of discrimination against women in the Gender treatment. Holding fixed beliefs about individual performances and average performance differences across the groups, we can ask whether women are any less likely to be hired when they are labeled as women rather than as even-month workers. Third, by replacing the gender labels in the Gender treatment with birth month labels in the Birth Month treatment, we are able to speak to the relevance of in-group preferences in hiring decisions, and whether such in-group preferences, are gender-specific.

We find that, on average, women in the Gender treatment are less likely to be hired than equally able men. That is, when presented with two workers with identical easy quiz performances, one of whom is a man and one of whom is a woman, employers are significantly less likely to hire the woman than the man on average. Perhaps surprisingly, however, this result is not specific to gender. When we turn attention to the Birth Month treatment, we observe similar levels of discrimination against even-month workers. This suggests that the behavior we observe in our stylized experiment is not driven by animus toward women and instead is more consistent with belief-based theories of discrimination. In particular, beliefs of average group differences are key.

We also find evidence of in-group preferences. In the Gender treatment, female employers hire

[5]This of course limits our ability to test the stereotype formation channel of Bordalo et al. (2016). We focus instead on decision-making given a certain set of beliefs.

women more often than male employers do. But, again, this result does not appear to be driven specifically by gender. When we consider the Birth Month treatment, we find similar in-group preferences: employers born in even months hire even-month employees more than odd-month employers do.

One important question is whether our results are driven by image concerns, such as not wanting to appear sexist. We take this concern seriously in our design, using additional questions to get at exactly this issue. We explore whether our results persist when there is a veil on employers' intentions, reducing image concerns. We operationalize this veil in a manner that is similar in spirit to the examination of risk as a veil in Exley (2015).[6] We continue to find that women are hired no less often when labeled as women than when labeled as even-month, suggesting that our failure to document clear evidence of taste-based discrimination against women is not driven simply by a fear of appearing sexist, or image concerns more generally. The introduction of a veil of intentions does, however, eliminate in-group preferences, suggesting a significant role for image concerns in driving these findings.

Our findings point to a key source of gender discrimination: beliefs. In this way, our work represents an advance within the empirical social science literature on identifying, measuring, and understanding discriminatory behavior. Our work also has important implications for practitioners interested in limiting discrimination within the workplace. First, it is important to recognize that we all hold beliefs about others based upon the groups to which these others belong (their gender, their race, their socioeconomic status, etc.). In many cases, these beliefs are inaccurate. Inaccurate beliefs are particularly likely when information is more scarce, and in contexts where stereotypes are prevalent. Second, the results from our experiment suggest that these beliefs may play a critical role in individual decision-making: beliefs may inform who we hire, who we promote, and who we turn to for leadership. When making decisions between a man and a woman, beliefs about average group-level differences across men and women may fuel discriminatory behavior even when the individual-level performance information on the man and woman is identical.

2 Design

Our studies are conducted on Amazon Mechanical Turk (MTurk). MTurk has a number of features that make it an appropriate and useful platform for our experiment. Most centrally, on Mturk, it is feasible to collect the large amount of data needed to be adequately powered for the statistical tests we run.[7] Importantly, previous work has shown that behavior among

[6] See also Dana, Weber and Kuang (2007), Haisley and Weber (2010), Danilov and Saccardo (2017), and for reviews, Kunda (1990) and Gino, Norton and Weber (2016).

[7] With 400 employers in each our two treatments, we have a statistical power of approximately 80% to detect a 10pp difference in hiring rates of women across our two treatments at the 5% level. Because we run several versions of our main hiring experiment, detailed below, in total we recruited 3,695 individuals to participate, making the study nearly impossible to run in a laboratory.

participants on MTurk and participants in laboratory studies is quite similar, including in the extent to which they pay attention to and follow instructions (Paolacci, Chandler and Ipeirotis, 2010; Halberda et al., 2012). Many classic laboratory studies have been successfully replicated on MTurk (Rand, 2012).

In a preliminary study, we collect performance information on four types of quizzes from participants: an easy and a hard math quiz, and an easy and a hard sports quiz.[8] We recruit 100 individuals for an advertised 40-minute academic study with a guaranteed completion fee of $6. Each of the four quizzes contained 10 multiple-choice questions. Workers had three minutes per quiz to answer as many questions as possible. To incentivize performance, workers received, as bonus payment within one week, 10 cents for each question they answered correctly, for one randomly-selected quiz. The workers were also aware that their performances would potentially be shown to other participants in follow-up experiments.

The purpose of this preliminary study is to produce a pool of available "workers" for hire in our hiring experiment. For each worker from this preliminary study, we have performance information that we can reveal to potential "employers," enabling us to ask employers to make incentivized hiring decisions over available workers.

In our main hiring experiment, we ask employers to make incentivized choices over available workers. We recruit 800 new MTurk participants for a 25-minute academic study with a guaranteed completion fee of $4. Employers receive information about the easy quiz performances of available workers. Then, they make a series of hiring decisions from different sets of available workers. Importantly, across every hiring decision they make, the employers have access to (at least some) information about the easy quiz performance of the available workers. To incentivize each hiring decision, the employer earns money based upon the hard quiz performance of the hired worker in each decision and knows that they will receive, as bonus payment within one week, any associated payoffs from one randomly-selected decision.

In this sense, our design mimics many real world hiring paradigms in three important respects. First, the employer receives some information about performance on an initial quality assessment (the easy quiz) for each candidate. Second, performance on this initial assessment (the easy quiz) serves as a signal of the candidate's capacity to perform well in the job (the hard quiz). Third, the employer stands to gain more by hiring workers who ultimately perform better on the job (the hard quiz).

We describe our experimental design, and treatment variations, in detail below. Appendix C provides full experiment instructions and screenshots of the experiment.

[8] We actually collect data on six total quizzes. The two extra quizzes are an additional easy math quiz and an additional easy sports quiz. We do not use performance on these two other quizzes in the hiring experiment, so we choose to omit them from our description of the design in order to reduce confusion.

2.1 Treatment variations

Our key treatment variation is whether the available workers are described to employers in terms of their gender, or not. Importantly, our goal is to hold everything else as fixed as possible among the workers across the two treatments, including employer beliefs about the performances of the groups of workers. To do this, we construct an admittedly strange, but very useful, set of available workers. Among the 100 workers who completed our preliminary performance study, we restrict our attention to two groups of workers: male workers born in odd months (who we will refer to as "male-odd-month" workers) and female workers born in even months (who we will refer to as "female-even-month" workers). These two groups of workers will be our available workers across both treatments of the study. All that will vary is how these workers are described to employers.[9]

While employers always make decisions between male-odd-month workers and female-even-month workers, employers view different labels for these groups of workers. Employers in the Birth Month treatment are not provided with the gender labels and see all information and decisions as between odd-month workers and even-month workers. Employers in the Gender treatment are not provided with the birth month labels and see all information and decisions as between male workers and female workers. The treatments are otherwise identical.

In addition, employers are randomly assigned to either the math or sports version of the study. In the math version of the study, employers see information about worker performance on the easy math quiz and are incentivized according to worker performance on the hard math quiz. In the sports version of the study, employers see information about worker performance on the easy sports quiz and are incentivized according to worker performance on the hard sports quiz. We conducted two versions of the study to better understand how the extent of discrimination in a male-typed domain might vary depending upon the seriousness or prestige of the task at hand. That is, we hypothesized that employers might feel that it is more acceptable or socially appropriate to discriminate against women in a sillier task, like sports, compared to a more educationally and career relevant task, like math. In most of the analysis below, we pool the two versions of the study together and discuss them jointly. In Section 3.3, we analyze the differences between the two versions.

2.2 Elicitation of Prior Beliefs

In the very first stage of the hiring experiment, we elicit employers' prior beliefs about the performance gap between male-odd-month workers and female-even-month workers. We explain

[9]This design involves the withholding of information from employers (worker gender in the Birth Month treatment, and worker birth month in the Gender treatment), but no mis-representation of information. Thus, we feel that it does not constitute deception. Notably, gender information (not to mention birth month information) is very often withheld from participants in experiments in which participants interact with each other — including in dictator games, trust games, bargaining games, and tournament competitions. Thus, while the withholding of this gender information is salient to the reader given our design, it is not particularly unusual in experimental paradigms.

that we have recruited a past set of MTurk workers to complete easy and hard 10-question quizzes in math (sports). Employers are asked to provide their beliefs about average group differences in performance on these quizzes. Employers in the Birth Month (Gender) treatment are asked about their prior beliefs about the average performance gap between odd-month (male) and even-month (female) workers. In particular, employers predict the difference in average scores between these groups of workers on a 10-question easy quiz and a 10-question hard quiz. Employers could indicate any feasible average difference (from -10 to 10 problems solved correctly), with positive numbers indicating a performance gap in favor of odd-month (male) workers in the Birth Month (Gender) treatment. These beliefs are not incentivized. We collect this information as a baseline before employers proceed to the Information Stage.

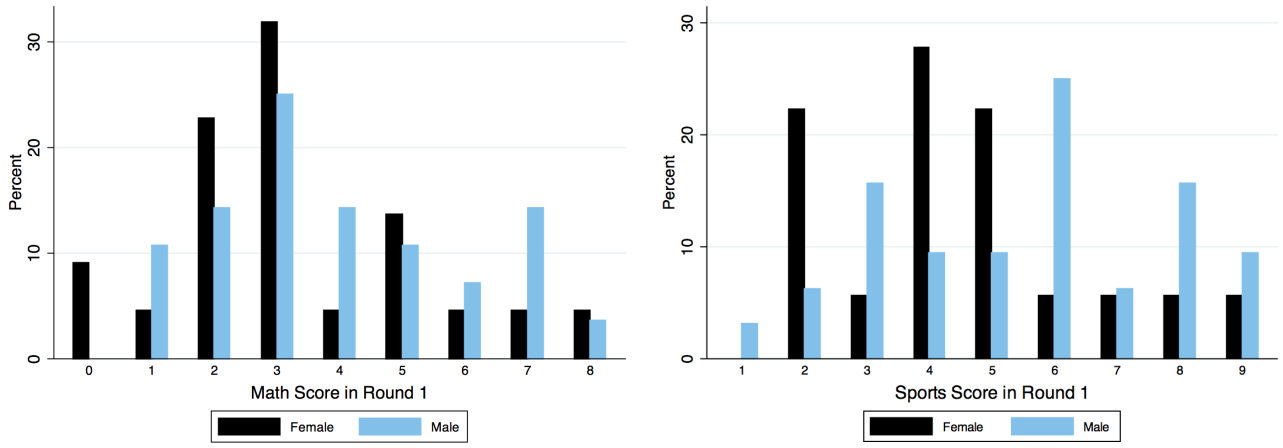
2.3 Information Stage

After providing prior beliefs, employers proceed to the Information Stage. The Information Stage is a critical component of our design. In order for the Birth Month treatment to serve as the right comparison treatment for the Gender treatment, employers must have similar beliefs about the performances of the two groups of workers in both treatments. That is, our goal is to have the perceived gender gap in performance be roughly equal to the perceived birth-month gap in performance when we compare across the two treatments. This is obviously unlikely in terms of prior beliefs. So, we achieve this by providing accurate and comprehensive information about performances on the easy quiz in our experiment. We show each employer the full distribution of performances on the easy quiz for female-even-month workers and male-odd-month workers. However, rather than just show this information once (i.e. present one histogram of performances), we reinforce this information by showing different subsamples of this data to the employer.

In particular, we draw various subsets of performances on the easy quiz of female-even-month workers and male-odd-month workers from our experiment. To give a concrete example, an employer sees the following subset of performances: a histogram of the performances of all female-even-month workers born between the between the 1^{st} - 15^{th} of the month compared to a histogram of the performances of all male-odd-month workers born between the 1^{st} - 15^{th} of the month. This allows the employer to compare performances among men and women (or among even-month and odd-month workers) among this particular subset. Then, they see a new subset of performances, selected by drawing workers from another range of birth dates. In total, employers see 12 different sets of distributions, each drawing from a different range of possible birth dates. The first 11 sets of distributions that employers see are formed by restricting the subset of workers to those born during different date ranges (see Appendix Table A.1 for more details on how these distributions are formed). The 11 distributions are unique but overlapping (that is, an employer only sees the screen with the distributions for workers born between the 1^{st} - 15^{th} of the month once, but she also sees a screen with the distributions for workers born

between the 1^{st} - 10^{th} days of the month). The order in which these 11 sets of distributions are shown is randomized at the participant level, and each is presented on a separate screen. For all participants, the 12^{th} and final set of distributions contains the full distributions, the histograms of the performances of all male-odd-month workers and female-even-month workers in our experiment.[10] Figure 1 illustrates how this information was displayed to participants, using the final set of full distributions.

Figure 1: Worker Performance on the Easy quiz



(a) Math Version

(b) Sports Version

Note: Panel A shows the performance distributions in the easy math quiz for female-even-month workers, who have an average performance of 3.27, and male-odd-month workers, who have an average performance of 3.96. Panel B shows the performance distributions in the easy sports quiz for female-even-month workers, who have an average performance of 4.50, and male-odd-month workers, who have an average performance of 5.50. The labels "Female" and "Male" shown here are those that would appear in the Gender treatment. In the Birth Month treatment, the labels would instead read "Even" and "Odd", respectively. Each employer saw these distributions as the final screen of the Information Stage.

In addition, we seek to increase the employers' engagement with these data by requiring them to make *incentivized* decisions based upon this information. For each set of the 12 sets of distributions they see, we ask employers to make an incentivized decision. For each set of distributions, the employers are asked to select one distribution from which they would like to hire. In the Birth Month treatment, the options are "The Even Month Group", "The Odd Month Group" and "Let Chance Determine the Group." In the Gender treatment, the options are "The Female Group", "The Male Group" and "Let Chance Determine the Group." If an employer

[10]We do not believe there is anything special about this particular construction of subsets of performance information. We felt it would be better for the employers to see this information in many stages, reinforcing the information, without simply repeating it verbatim. Dividing by day of birth was simply a convenient and available formulation for our design.

selects "Let Chance Determine the Group," the computer flips a coin to determine which group is hired from. We provide the "Let Chance Determine the Group" option in order to better identify indifference among our population; with this option available, it seems more likely that a choice of one of the other two groups reflects a strict preference.

If one of these decisions is randomly selected as the decision that counts for payment, one worker from the selected group is chosen at random to be hired. The hired worker receives an additional 25 cents as bonus payment. Employers receive 10 cents for each question answered correctly by the hired worked on the hard quiz.

The main goal of these choices is to encourage employers to engage with this information, incorporating it into their beliefs about the average performance gap. While we can consider choices made during this stage of the experiment when evaluating the degree of discrimination, they are not as informative about the drivers of discrimination. This is because beliefs are likely not fixed at this stage in the experiment, and may evolve differently across the two treatments. Nonetheless, we discuss these Information Stage hiring decisions, and what we can learn from them, in Section 3.5.

2.4 Elicitation of Posterior Beliefs

After completing the Information Stage, employers are re-asked all of the beliefs questions from the Elicitation of Prior Beliefs stage. That is, they see the same beliefs questions again so that we can collect posterior beliefs. Again, these beliefs are not incentivized. The key data for us are employers' posterior beliefs of average differences in performance across the two groups. Our goal is to ensure similar beliefs of average performance differences across the two treatments. That way, we can isolate the role for gender-specific taste considerations, holding beliefs fixed, by comparing across the Gender and Birth Month treatment.

2.5 Hiring Stage

After the Information Stage and Elicitation of Posterior Beliefs, we ask employers to make a series of hiring decisions between specific pairs of workers. In each hiring decision, employers in the Birth Month treatment are asked whether they want to hire "The Even-Month Worker," the "Odd-Month Worker," or "Let Chance Determine Who is Hired." Employers in the Gender treatment are asked whether they want to hire the "The Female Worker," "The Male Worker," or "Let Chance Determine Who is Hired." Again, the chance option allows employers to explicitly express indifference. For each hiring decision, employers learn the two workers' exact performances on the easy quiz. We vary the performances of the available workers and the payoffs to hiring each across a series of 54 decisions, split across six screens (within-subject).

On each screen in the Hiring Stage, employers make nine hiring decisions. For each, they are told the easy quiz performances of the two workers in the pair. In three of the hiring decisions, the available workers have the same performance on the easy quiz. The number of questions answered correctly (out of 10) by the male-odd-month worker versus the female-even-month

worker is (i) 4 versus 4, (ii) 6 versus 6, and (iii) 8 versus 8. In the other six hiring decisions on the screen, the male-odd-month worker has a weaker performance than the female-even-month worker: (iv) 4 versus 5, (v) 4 versus 6, (vi) 4 versus 7, (vii) 4 versus 8, (viii) 6 versus 7, and (ix) 6 versus 8. We choose to focus on decisions in which the female-even-month worker weakly outperforms the male-odd-month worker so that we can easily classify all decisions not to hire the female worker as discrimination against women.

Across the screens in the Hiring Stage, we vary the payoffs to hiring each worker. On the first Hiring Stage screen, the payoff to hiring either worker is 10 cents for each question answered correctly by their hired worker on the hard quiz.

In these stark, side-by-side hiring decisions, employers may feel that choosing not to hire the female-even-month worker (who always has a weakly better performance) is harmful to their self-image or social-image. This might be particularly true in the Gender treatment due to concerns about perceived sexism. A key question we wish to ask is whether social desirability concerns, or image concerns more generally, limits the extent of discrimination against the female-even-month worker. To do this, we introduce additional sets of hiring decisions where we provide a clear and easily justifiable "excuse" for the employer not to hire the female-even-month worker, similar in spirit to Exley (2015).[11]

Paralleling Exley (2015), the potential excuse we provide takes the form of risk. We present again the hiring decisions from the first screen of the hiring stage. But now, we alter each decision so that the decision to hire a female-even-month worker becomes risky. In these decisions, the payoff remains 10 cents per correct answer on the hard quiz *if* the employer hires a male-odd-month worker. But, if the employer hires a female-even-month worker, there is some risk that the employer will receive nothing. In particular, if an employer hires a female-even-month worker, she receives 10 cents for each question correctly answered by that worker on the hard quiz with P% chance but no payment for that particular hiring decision with (100-P)% chance. Across the second through sixth Hiring Stage screens, the risk of hiring the female-even-month worker increases, as we decrease P from 99, 95, 90, 75 to 50.

From an image perspective, the introduction of this risk can help to serve as a veil on employer intentions. If an employer does not hire a woman when the payoffs to hiring either the man or the woman are the same, this could easily be interpreted as discriminatory or sexist behavior. But, if that same employer does not hire the woman under the risky decisions, an alternative interpretation is more readily available: the employer is just profit-maximizing. In this way, employers who would like to discriminate against women in riskless decisions — but feel reluctant to do so due to image concerns — may feel more able to do so in the risky decisions. Of course,

[11] Danilov and Saccardo (2017) use a related approach in their paper on ethnic discrimination. While they find no evidence of discrimination by Germans toward Turks in a basic dictator game, when they require that the sender tell a lie in order to provide the receiver with the generous outcome, Germans are less likely to do this for a Turk than a German.

this explanation seems most relevant in the Gender treatment (where image concerns about sexism are clear). So, our Birth Month control condition is helpful in isolating this mechanism.

The introduction of risk decreases the expected payment from hiring a female-even-month worker in *both treatments*. So, we expect less hiring of female-even-month workers in the risky decisions than in the riskless decisions under both treatments. But, if in addition, we see a larger employer reaction to the introduction of risk in the Gender treatment than in the Birth Month treatment (that is, a bigger drop in the rate of hiring female-even-month workers), this suggests that some of the reaction in the Gender treatment is not just about the expected decrease in payoff. Instead, it may also suggest an enabling of discrimination that had previously been mitigated by image concerns in the riskless decisions.

If one of these 54 hiring decisions is randomly selected to count for payment, employers receive the amount of money earned in the selected decision as an additional bonus payment. Their hired worker from the selected decision also receives 25 cents as additional bonus payment.

This design is summarized in Figures 2 and 3.

Figure 2: Preliminary Study to Create Pool of Workers

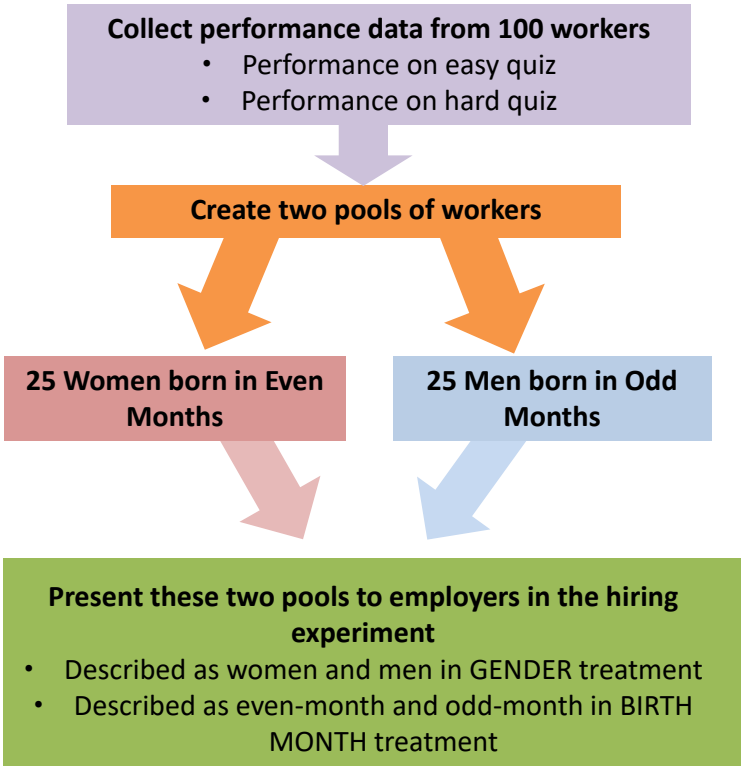
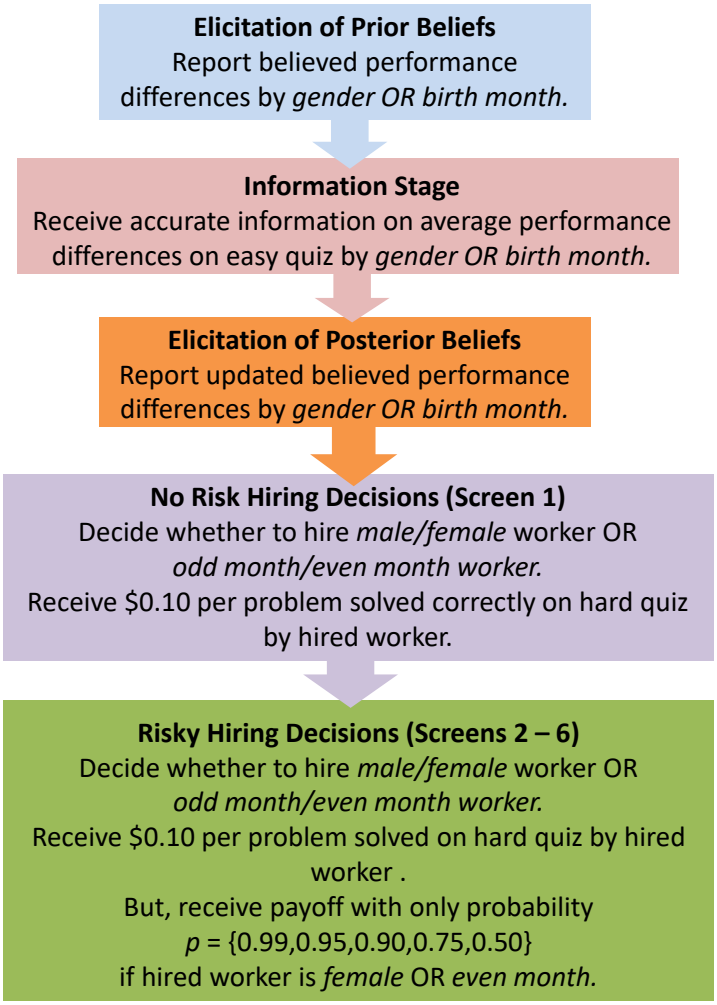


Figure 3: Overview of Hiring Experiment



3 Results

We begin by examining employers' beliefs in Section 3.1. We then investigate hiring decisions in Section 3.2, revealing an important role for beliefs. To provide more insight, we explore how the extent of discrimination varies by domain (math versus sports) in Section 3.3, and how the extent of discrimination varies by type of employer (male versus female) in Section 3.4.

We start first by considering decisions from the Hiring Stage where employers can choose between a female-even-month worker and a male-odd-month worker with identical easy round performances. We focus on these decisions for two reasons. First, because the workers have identical easy round performances, it is a clear-cut environment in which to look for discrimination: absent discrimination, we expect female-even-month and male-odd-month workers to be hired at identical rates. Second, since these decisions occur after the Information Stage where all group-level information is provided, we can more reasonably assume that beliefs have been fixed and are

similar across treatments. This allows us not just to document the extent of discrimination, but also to speak to its sources. After this main analysis, we explore all other decisions that employers make during our study. In particular, Section 3.5 details the hiring decisions that employers make during the Information Stage as well as the hiring decisions that employers make when the female-even-month worker has a better easy round performance than the male-odd-month worker in the Hiring Stage. Finally, we document the robustness of our findings by showing that employer decisions look quite similar in a highly simplified follow-up hiring experiment with a new sample of employers.

3.1 What beliefs do employers hold?

Figure 4 displays the kernel densities of employers' prior beliefs about the performance gap on the hard and easy quizzes between male-odd-month workers and female-even-month workers. As we expect, employers initially believe that male workers outperform female workers by a significantly larger amount than odd-month workers outperform even-month workers (see Panels A and C; the p-value for across-treatment differences in prior beliefs is <0.01 for both the easy and hard quizzes, using either a K-S test or a t-test). After employers complete the Information Stage, however, posterior beliefs about these gaps are no longer significantly different across treatments (see Panel B and D; the p-value for across-treatment differences in posterior beliefs is >0.10 for both the easy and hard quizzes, using either a K-S test or a t-test).[12]

Our Information Stage "works" in the sense that, afterwards, the extent to which employers in the Gender treatment believe male workers outperform female workers is the same as the extent to which employers in the Birth Month treatment believe odd-month workers outperform even-month workers. Given that, on average, employers believe there is a significant gap in average performance, this provides a clear motive to statistically discriminate. And, since beliefs are similar across treatments, we should expect a similar degree of this type of belief-based discrimination across treatments. Thus, if we see different levels of discrimination across treatment, we can attribute this to gender-specific tastes, rather than beliefs.

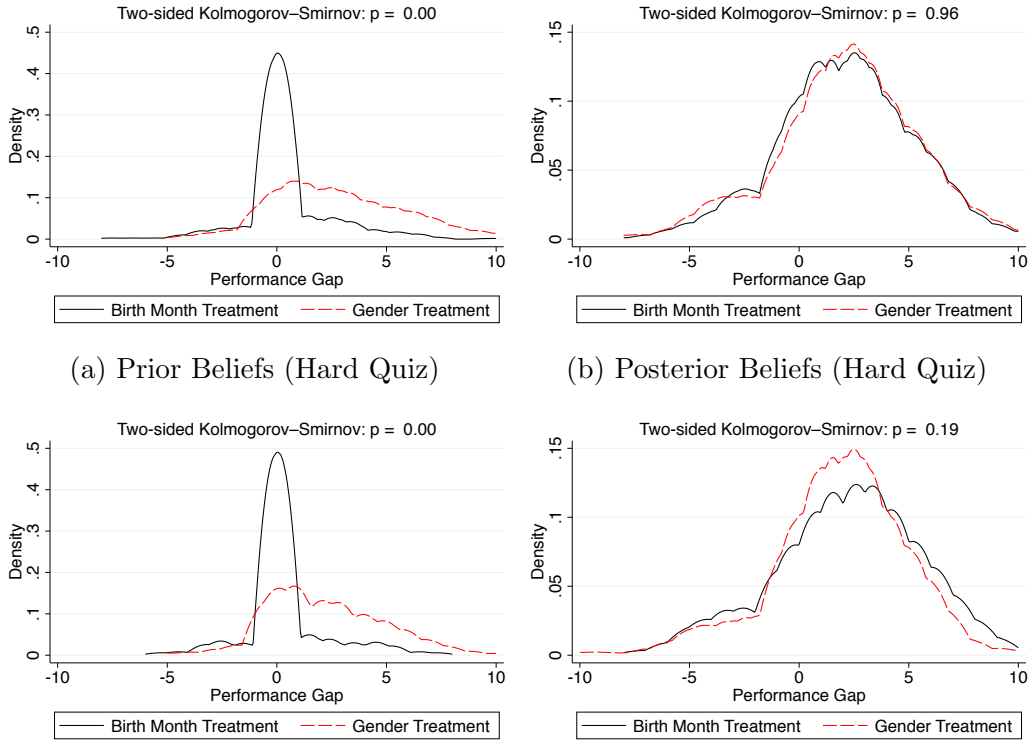
Although not central to out investigation, we can also speak to the accuracy of these beliefs. We note that posterior beliefs get the direction of the difference right on average, but exaggerate it — believing male-odd-month workers outperform female-even-month workers on both the hard and easy quizzes by more than they actually do.[13] This is in line with the model of biased

[12] Appendix Table A.2 confirms that this belief convergence is on average statistically significant and shows that it is similar when separately considering workers' performance on sports or math.

[13] In the sports version, the average posterior belief of this gap is 2.76 on the hard quiz and 2.62 on the easy quiz, while the actual average gap in performance is 1.1 on the hard quiz and 1.0 on the easy quiz. If we test whether the believed gap is significantly different than the observed gap in performance, we reject the null of equality with $p<0.01$ for both the hard and easy quiz. Similarly, in the math version, the average posterior belief of the gap is 1.54 on the hard quiz and 1.07 on the easy quiz, while the actual average gap is 0.451 on the hard quiz and 0.692 on the easy quiz. Again, believed gaps are significantly different than observed gaps, with $p<0.01$ on the hard quiz and $p = 0.053$ on the easy quiz.

beliefs and stereotyping by Bordalo et al. (2016) and the evidence presented in Bordalo et al. (2019). It is also worth noting that male-odd-month workers directionally outperform female-even-month workers on the hard quiz even when controlling for easy quiz performance.[14] Thus, an employer who statistically discriminates — i.e. practices belief-based discrimination based upon accurate beliefs — would hire male-odd-month workers over female-even-month workers when they have the same easy quiz performance. Of course, a decision-maker who holds biased beliefs — exaggerating the gap — would similarly make this hiring decision. In our investigation, we will focus on the predictive power of beliefs generally (both accurate and inaccurate), and focus on distinguishing any belief-based form of discrimination from taste-based considerations.

Figure 4: Beliefs about performance gap between male-odd-month workers and female-even-months workers



(a) Prior Beliefs (Hard Quiz)

(b) Posterior Beliefs (Hard Quiz)

(c) Prior Beliefs (Easy Quiz)

(d) Posterior Beliefs (Easy Quiz)

3.2 Is there discrimination against women?

We begin by considering hiring decisions between a male-odd-month worker and a female-even-month worker when all else is equal.[15] Employers learn that the payoff rule from hiring

[14] In particular, a regression of the hard quiz performance on an indicator for female-even-month workers and indicators for each easy round performance level indicates that male-even-month workers, relative to female-even-month workers, answer an average of 0.422 more questions correctly on the hard sports quiz ($p = 0.464$) and an average of 0.538 more questions correctly on the hard math quiz ($p = 0.089$).

[15] For raw data and summary information on employer choices over all decisions in the experiment, see Appendix Tables A.3, A.4. and A.5.

either worker is the same: they receive 10 cents per correct answer on the hard quiz by the hired worker. Employers also receive identical easy quiz performance information on both workers: both workers are known to answer 4, 6, or 8 (out of 10) questions correctly on the easy quiz.[16]

If employers neither engage in belief-based nor taste-based discrimination, we expect that they should hire female-even-month workers 50% of the time.[17] This is not the case. Employers in the Gender treatment only hire female workers in 43% of decisions, significantly below the 50% benchmark.

This discrimination is not specific to gender, however. In the Birth Month treatment, an even-month worker is hired in just 37% of decisions, also significantly below the 50% benchmark. Employers in both treatments discriminate against workers associated with the lower-performing group, consistent with beliefs driving discrimination against women, rather than tastes.

Table 1 provides corresponding regression results, when clustering standard errors at the employer-level, from specifications of the following form:

$$\begin{aligned} \textit{P(female-even-month worker is hired)} &= \beta_1 \textit{Gender Treatment} \\ &\quad [+\beta_2 \textit{Posterior(easy gap)}] \\ &\quad [+\beta_3 \textit{Posterior(hard-easy gap)}] \\ &\quad + \sum_{i=1}^{2} \sum_{j=1}^{3} v_i \times d_j + \epsilon \end{aligned} \tag{1}$$

where *P(female-even-month worker is hired)* is the probability with which a female-even-month worker is hired, β_1 captures the impact of being in the Gender Treatment, β_2 and β_3 capture the coefficient estimates on controls for beliefs (see the table note of Table 1 for details), and v_i are dummies for the math versus sports version while d_j are dummies for the performance levels of the workers.

Column 1 of Table 1 Panel A confirms that taste-based considerations, if anything, work in favor of women: female-even-month workers are 6 percentage points more likely to be hired when they are labeled as female workers in the Gender treatment than when they are labeled as even-month workers in the Birth Month treatment.

In Column 2 of Table 1 Panel A, we directly explore the role of beliefs. We add to the regression two belief measures: the employer's posterior belief of the average performance gap in the easy quiz, and her imputed belief of "differential improvement" from the easy to the hard

[16] In the analysis below, we pool these decisions (4 v 4, 6 v 6, and 8 v 8). Results are quite similar for each of the three decisions independently; see Appendix Table A.6.

[17] We code the probability of a female-even-month worker being hired as 1 if she is hired with certainty, 0.5 if chance determines who is hired, and 0 if a male-odd-month worker is instead hired with certainty. Appendix Table A.7 shows that our main results replicate if we consider other methods of classifying decisions across these three options.

quiz.[18] Both measures are signed so that a positive gap indicates a belief of a male-odd-month advantage. Column 2 shows that beliefs are highly predictive of decisions: employers who believe the performance gap is larger are less likely to hire the worker from the lower-performing group.[19]

Columns 3 - 4 of Table 1 Panel A present results from when a "veil" for employers' intentions is provided. We do this via the addition of risk into the payoff from hiring a female-even-month worker but not a male-odd-month worker. The key result is that female-even-month workers are now hired at equal rates across the two treatments. The fact that, under risk, women are now no more likely to be hired in the Gender treatment than in the Birth Month treatment suggests that some of the female preference we observe in the stark, riskless environment may be driven by image concerns.[20] Indeed, the odd columns in Appendix Table A.10 show we do not observe evidence for female preference (greater hiring of women in the Gender treatment than in the Birth Month treatment) when separately examining our results for each level of risk. This is true even when the risk is only a 1% chance of no payment from hiring the female-even-month worker.

But, it is worth emphasizing here that women are still *no less likely* to be hired in the Gender treatment than in the Birth Month treatment, even under risk. Thus, gender-based animus against women does not arise even when a veil for employers' intentions exists. It does not appear to be the case that the absence of taste-based discrimination against women in riskless decisions was only an artifact of image concerns or social desirability bias. Even when intentions are veiled, women are no worse off being labeled as women than being labeled as even-month.

Taken together, our results may be summarized as follows. When women are the lower-performing group, they are discriminated against even when their individual performances are as good as their counterparts. This type of discrimination, however, is not specific to their gender. Rather, it appears driven by beliefs. Female workers are treated no worse than even-month workers, even when there is a veil for employers' intentions.

After seeing these results, we were curious about the extent to which the observed discrimination against women was driven by the fact that participants receive ample, detailed information about the average gender gap in performance in the easy quiz. We hypothesized that this information may have helped participants feel justified in their decision to discriminate (either through

[18]Differential improvement is the idea that participants may believe that the easy quiz performance is differentially predictive of the hard quiz performance across the two groups. To measure this, we difference the participant's belief of the hard quiz performance gap and the easy quiz performance gap, and include this difference-in-differences as a control. By including both differential improvement and posterior belief of the easy quiz gap, we allow our specifications to depend on beliefs about both the easy and hard quiz. Our results are not sensitive to this particular construction. Results available upon request.

[19]Column 1 of Appendix Table A.8 shows that beliefs are not more predictive in the Gender treatment. So, while we cannot directly rule out that beliefs in the Gender treatment are more or less strongly held than beliefs in the Birth Month treatment, it is not the case that one set of beliefs is more predictive than the other.

[20]While the preference for hiring women relative to even-month is insignificant under risk, we cannot reject that the effects are the same in the risk and riskless decisions. See Columns 1 and 2 of Appendix Table A.9.

Table 1: Hiring decisions between workers with the same performance

	Same payoff rule		Risk in payoff rule only for female-even-month workers	
	(1)	(2)	(3)	(4)
Panel A: Gender and Birth Month Treatments				
Gender Treatment	0.061***	0.057***	0.026	0.024
	(0.019)	(0.018)	(0.019)	(0.018)
Posterior(easy gap)		-0.023***		-0.018***
		(0.003)		(0.003)
Posterior(hard-easy gap)		-0.011**		-0.011**
		(0.005)		(0.004)
Decision FEs	yes	yes	yes	yes
Observations	2400	2400	12000	12000
Panel B: Gender and Gender-No-Information Treatments				
No Information	0.002	0.004	0.005	0.008
	(0.020)	(0.019)	(0.019)	(0.019)
Posterior(easy gap)		-0.021***		-0.015***
		(0.004)		(0.004)
Posterior(hard-easy gap)		-0.014***		-0.015***
		(0.005)		(0.004)
Decision FEs	yes	yes	yes	yes
Observations	2409	2409	12045	12045

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors are clustered at the employer-level and shown in parentheses. The results are from ordinary least squares regressions of the probability with which a female-even-month worker is hired over a male-odd-month worker, among employers making hiring decisions between pairs of workers who have the same performances. ***Gender Treatment*** is an indicator for employers in the Gender treatment. ***Posterior(easy gap)*** is an employer's posterior belief about the average performance gap in the easy quiz. ***Posterior(hard − easy gap)*** is an employer's posterior belief about the average performance gap in the hard quiz minus the employer's posterior belief about the average performance gap in the easy quiz. ***No Information*** is an indicator for employers in the *No Information Stage* version. Decisions FEs include 6 fixed effects (with one excluded) that capture whether hiring decisions are between workers with equal performances on the sports quiz (4 v 4, 6 v 6, or 8 v 8) or between workers with equal performances on the math quiz (4 v 4, 6 v 6, or 8 v 8). Columns 1 - 2 involve hiring decisions where the payoff rule is always the same, while Columns 3 - 4 involve hiring decisions where the payoff rule is risky when hiring female-even-month workers and riskless when hiring male-odd-month workers. Panel A involves hiring decisions from the Birth Month and Gender treatments, while Panel B involves hiring decisions from the Gender and Gender-No-Information treatments.

experimenter demand, or because their beliefs are so well-informed). Would participants be willing to discriminate without any information on the distributions of easy quiz performances?

To test this hypothesis, we ran a new *No Information Stage* version of the experiment that simply removed the Information Stage for employers. All other aspects of the design were kept the same, and we again recruited 800 individuals from MTurk. These results are presented in Panel B of Table 1, both without risk (Columns 1 - 2) and with risk (Columns 3 - 4). We ask whether

female workers are hired as often in the Gender treatment of the *No Information Stage* version as in our baseline Gender treatment. Column 1 shows that the level of discrimination against women is identical without the Information Stage; this remains true when we control for beliefs in Column 2, and when we consider the risky decisions in Columns 3 and 4. Providing ample information about differences by gender neither promotes nor inhibits discrimination against women above and beyond what we find when employers only act on their priors. Employers are willing to discriminate even based on arguably less informed beliefs.[21]

3.3 Does the domain matter?

Employers in our experiment are randomized into one of two domain conditions: a math condition or a sports condition. Both math and sports have been perceived as domains of male advantage by past experimental participants who completed similar quizzes (Coffman, 2014; Bordalo et al., 2019). In this sense, we would expect belief-based discrimination against women in both domains. However, consistent with preliminary investigations by Bordalo et al. (2019), we hypothesized that employers may feel more at ease discriminating in a domain like sports as compared to math, as it is arguably less academically relevant, less prestigious, and less ego relevant. Put differently, the image concerns associated with discriminating against women in math might be much stronger than those associated with discriminating against women in sports.

We pre-tested this hypothesis empirically among the workers who completed our preliminary study. After completing all of the quizzes, participants in the preliminary study were asked a question about the social appropriateness of openly expressing beliefs about performance differences between men and women, in either math or in sports. In particular, they are told: imagine that some individual held the following belief: on average, it is likely that men performed better than women in the [math/sports] questions. They are then asked: how socially appropriate would it be for that individual to share that belief with others? They could select, very socially inappropriate (1), somewhat socially inappropriate (2), somewhat socially appropriate (3), or very socially appropriate (4). Following Krupka and Weber (2013), we incentivize these beliefs by providing a bonus payment if participants provide the modal answer provided by other participants. Thus, these answers can be interpreted as participants' beliefs about the norms surrounding the appropriateness of beliefs of gender differences in these domains.

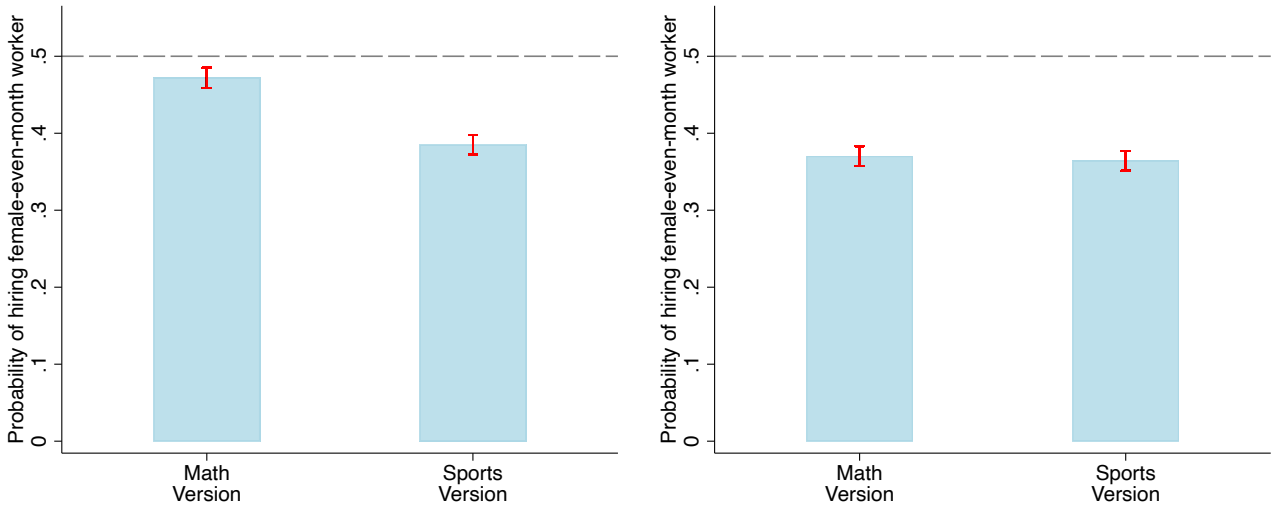
Among our sample of 100 performance pool participants, we see significant differences in these assessments for math versus sports. The average rating for sports is 3.15, somewhere between somewhat socially appropriate and very socially appropriate. The average rating for math, on the other hand, is 2.35, closer to somewhat socially inappropriate ($p<0.001$ for the across-domain

[21] We present a replication of our main beliefs figure, Figure 4, for the *No Information Stage* version as Appendix Figure A.2. While employers in the Gender treatment still believe there is a significant male advantage in performance absent information, not surprisingly, there is no significant believed difference in performance in the Birth Month treatment absent our information. Quite understandably then, discrimination against even-month workers does not persist in the *No Information Stage* version (see Appendix Figure A.1).

difference). This is consistent with our hypothesis that image concerns related to discriminating against women in math may be more severe than those associated with discriminating in sports.

Taking this hypothesis to our hiring experiment, we predict that individuals will be more reluctant to discriminate against women in math than in sports in our Gender treatment, particularly without the veil of risk. We expect no differences across math and sports in the Birth Month treatment. In Figure 5, we show the average rate at which the female-even-month worker is hired in riskless decisions, when both workers have the same performance. We split the data both by treatment (Gender versus Birth Month) and also by domain (math versus sports). There is, on average, discrimination against the female-even-month worker across both domains and treatments. But, consistent with our hypothesis, we see significantly more discrimination against the female-even-month worker in sports than in math within the Gender treatment. On the other hand, and again consistent with our hypothesis, the level of discrimination against the female-even-month worker in the Birth Month treatment is very similar across sports and math.

Figure 5: Math vs Sports Version: Hiring decisions between workers with same performance and payoff rule



(a) Gender Treatment

(b) Birth Month Treatment

We formalize this analysis in Table 2, where we reproduce Table 1 separately for math and sports. Columns (1) and (2) consider riskless decisions. In math, we see a strong impact of the Gender treatment, with the female-even-month worker being significantly more likely to be hired when she is labeled as female rather than even-month. This points to considerable taste-based preferences in favor of women in math, even conditional on stated beliefs (Column 2). In sports, it is still the case that women are, if anything, more likely to be hired under the Gender rather than the Brith Month treatment, but the effect is much smaller and is statistically insignificant. This suggests that there is much less "affirmative action" in sports than in math.

Table 2: Comparing the math and sports versions: hiring decisions between workers with the same performance

	Same payoff rule		Risk in payoff rule only for female-even-month workers	
	(1)	(2)	(3)	(4)
Panel A: Math Version				
Gender Treatment	0.102***	0.084***	0.071***	0.061**
	(0.027)	(0.026)	(0.027)	(0.027)
Posterior(easy gap)		-0.028***		-0.020***
		(0.005)		(0.005)
Posterior(hard-easy gap)		-0.015*		-0.016**
		(0.008)		(0.008)
Decision FEs	yes	yes	yes	yes
Observations	1200	1200	6000	6000
Panel B: Sports Version				
Gender Treatment	0.021	0.026	-0.019	-0.014
	(0.026)	(0.026)	(0.025)	(0.025)
Posterior(easy gap)		-0.018***		-0.016***
		(0.004)		(0.005)
Posterior(hard-easy gap)		-0.008		-0.008
		(0.005)		(0.005)
Decision FEs	yes	yes	yes	yes
Observations	1200	1200	6000	6000

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors are clustered at the employer-level and shown in parentheses. The results are from ordinary least squares regressions of the probability with which a female-even-month worker is hired over a male-odd-month worker, among employers making hiring decisions between pairs of workers who have the same performances. ***Gender Treatment*** is an indicator for employers in the Gender treatment. ***Posterior(easy gap)*** is an employer's posterior belief about the average performance gap in the easy quiz. ***Posterior(hard − easy gap)*** is an employer's posterior belief about the average performance gap in the hard quiz minus the employer's posterior belief about the average performance gap in the easy quiz. Decisions FEs include 3 fixed effects (with one excluded) that capture whether hiring decisions are between workers with equal performances of 4 v 4, 6 v 6, or 8 v 8. Columns 1 - 2 involve hiring decisions where the payoff rule is always the same, while Columns 3 - 4 involve hiring decisions where the payoff rule is risky when hiring female-even-month workers and riskless when hiring male-odd-month workers. Panel A involves hiring decisions from the math version, while Panel B involves hiring decisions from the sports version.

When looking at our pooled analysis, we concluded that female-even-month workers were significantly less likely to be hired than male-odd-month workers on average, conditional on having identical performances. And, this discrimination was driven by beliefs, not gender-specific tastes: female-even-month workers were no less likely to be hired when labeled as female than when labeled as even-month. Splitting the analysis by math and sports reinforces these findings,

as these results replicate within both individual domains. The across-domain differences that do emerge provide additional insights into our results. The taste-based preference that seems to work in favor of women (revealed by the Gender versus Birth Month comparison) is much more prominent in math than in sports, consistent with the idea that employers may be more reluctant to discriminate against women in areas where there are stronger social norms against it. Just as employers' behavior seems to reveal that they think it is more "okay" to discriminate against even-month workers than female workers, their behavior also seems to reveal that is it more "okay" to discriminate against women in sports than in math.

In Columns (3) and (4), we consider the risky decisions, again parelleling the analysis of Table 1. Across both math and sports, the introduction of risk directionally reduces the size of the Gender treatment effect, but not by a significant amount. In math, even with the introduction of risk, women are still significantly more likely to be hired in the Gender treatment than in the Birth Month treatment. In sports, there continues to be no significant differences across the Gender and Birth Month treatments.

3.4 Does the gender of the employer matter?

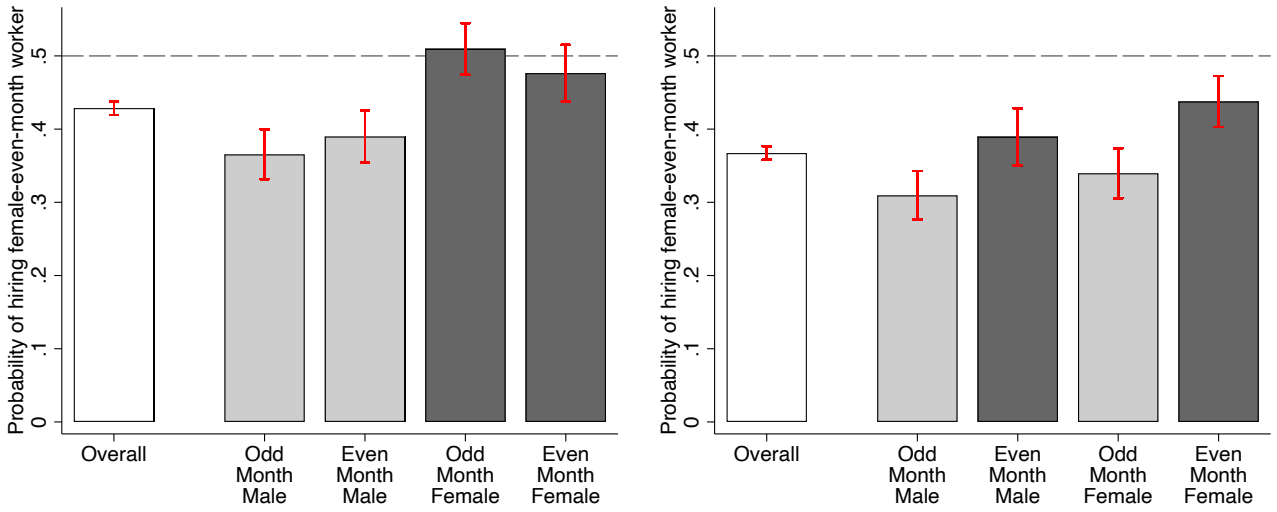
Figure 6 shows that there is substantial heterogeneity in the extent of discrimination when considering the four types of employers in our experiment as defined by their gender and their birth month. In the Gender treatment, male employers engage in substantial discrimination against female workers, hiring them less than 40% of the time. By contrast, and consistent instead with no discrimination, female employers hire female workers approximately 50% of the time.[22] Thus, in our setting, when gender is known, men discriminate against women, while women do not seem to discriminate against either women or men.

But, what drives this pattern? Is this something specific to gender? Our Birth Month treatment reveals that this result appears to be driven by in-group preferences. Just as men hire women less often than women do, odd-month employers hire even-month workers less often than even-month employers do. While odd-month employers hire even-month workers 32% of the time, even-month employers hire even-month workers 41% of the time. Thus, across both treatments, there is less discrimination by in-group employers (female employers in the Gender treatment, even-month employers in the Birth Month treatment).

In other words, it is not simply that female employers discriminate less often than male employers against women. Instead, employers discriminate less often against "in-group" members more broadly. This is consistent with previous work on in-group preferences that documents that shared social identities can impact economic outcomes, including studies using both minimal group paradigms and natural identities (see Tajfel et al. (1971), Brewer (1979), Chen and Li

[22]Past evidence on whether female employers are in general more likely to hire female workers is mixed, see for instance Bagues and Esteve-Volart (2010) and De Paola and Scoppa (2015). In recent work, Cappelen, Falch and Tungodden (2019) find that women are more likely to discriminate against low-performing males than men are.

Figure 6: Hiring decisions between workers with same performance and payoff rule



(a) Gender Treatment

(b) Birth Month Treatment

(2009), Chen and Chen (2011), and Chen et al. (2014)).

We explore this pattern in Table 3. We follow exactly the specifications of Table 1, but start by replacing the treatment dummy with an in-group dummy (even-month employer in the Birth Month treatment, or female employer in the Gender treatment) in Column 1. In Columns 2 and 3, we add back in our main treatment dummy and interact it with the in-group dummy, to ask whether the extent of in-group preferences varies by treatment. Column 1 documents statistically significant in-group preferences, while Column 2 shows that the extent of in-group preferences does not vary across treatment. In-group preferences are no stronger in the Gender treatment than in the Birth Month treatment.

One could ask whether these in-group preferences are well-explained by differences in beliefs by type of employer. In-group status does appear to influence posterior beliefs. In particular, within the Gender treatment, male employers on average believe the male advantage in performance on the easy quiz is 2.39 points on average, while women believe it is just 1.16 points on average. Similarly, within the Birth Month treatment, odd-month employers believe the odd-month advantage in performance on the easy quiz is 2.48 points on average, while even-month employers believe it is 1.91 points on average. However, these differences in beliefs do not explain the extent of in-group preferences we observe in our hiring decisions. This can be seen in Column 3, where we add the full set of employer beliefs to the regression model. While beliefs are once again strongly predictive of employer decisions, we see significant in-group preferences on top of stated beliefs.

Interestingly, however, in our setting, in-group preferences appear quite malleable. When provided with a veil on their intentions via the introduction of risk, employers do not demonstrate

Table 3: Hiring decisions between workers with the same performance

	Same payoff rule			Risk in payoff rule only for female-even-month workers		
	(1)	(2)	(3)	(4)	(5)	(6)
In-Group	0.098***	0.090***	0.081***	0.012	0.008	0.001
	(0.019)	(0.026)	(0.025)	(0.019)	(0.025)	(0.024)
Gender Treatment		0.055**	0.057**		0.022	0.025
		(0.025)	(0.025)		(0.025)	(0.025)
*Gender Treatment*In-Group*		0.020	0.007		0.009	-0.002
		(0.037)	(0.036)		(0.038)	(0.037)
Posterior(easy gap)			-0.021***			-0.018***
			(0.003)			(0.003)
Posterior(hard-easy gap)			-0.011**			-0.011**
			(0.005)			(0.004)
Decision FEs	yes	yes	yes	yes	yes	yes
Observations	2400	2400	2400	12000	12000	12000

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors are clustered at the employer-level and shown in parentheses. The results are from ordinary least squares regressions of the probability with which a female-even-month worker is hired over a male-odd-month worker, among employers making hiring decisions between pairs of workers who have the same performances. ***In-Group*** is an indicator for employers who know they share some demographic characteristic with the female-even-month workers (i.e., even-month employers in the Birth Month treatment and female employers in the Gender treatment). ***Gender Treatment*** is an indicator for employers in the Gender treatment. ***Posterior(easy gap)*** is an employer's posterior belief about the average performance gap in the easy quiz. ***Posterior(hard − easy gap)*** is an employer's posterior belief about the average performance gap in the hard quiz minus the employer's posterior belief about the average performance gap in the easy quiz. Decisions FEs include 6 fixed effects (with one excluded) that capture whether hiring decisions are between workers with equal performances on the sports quiz (4 v 4, 6 v 6, or 8 v 8) or between workers with equal performances on the math quiz (4 v 4, 6 v 6, or 8 v 8). Columns 1 - 3 involve hiring decisions where the payoff rule is always the same, while Columns 4 - 6 involve hiring decisions where the payoff rule is risky when hiring female-even-month workers and riskless when hiring male-odd-month workers. All hiring decisions are from the Birth Month and Gender treatments.

in-group preferences (see Columns 4 - 6 of Table 2, and the odd columns in Appendix Table A.10 to examine the results separately for each level of risk).[23] This suggests that the in-group preferences in the stark environment of the riskless decisions is likely reflective of image concerns, including potentially experimenter demand. Alternatively, it may be that in an environment with limited reasons for selecting one worker over another (i.e. the two workers have identical easy-round performances), in-group preferences serve as a "tie-breaker" of sorts. But, with the introduction of risk, this tie-breaker is less relevant, significantly reducing the role of in-group

[23]See Columns 3 and 4 of Appendix Table A.9 for an interacted model.

preferences.[24]

3.5 Do our results extend to other decisions?

In Sections 3.2, 3.3, and 3.4, we focus on hiring decisions between two workers with identical easy round performances. It is in these decisions that we see clear evidence of discrimination: differential treatment by employers of two employees with identical performance information. In this section, we examine whether our results extend to additional hiring decisions where: (i) the individual performances of the two workers is uncertain, or (ii) the female-even-month worker has a better easy round performance than the male-odd-month worker does.

The first set of additional hiring decisions involves the 12 decisions that employers make as part of the Information Stage. In each of these decisions, the distribution of easy round performances for some subset of female-even-month workers and the distribution of easy round performances for some subset of male-odd-month workers are displayed. Employers then choose between a female-even-month worker and male-odd-month worker, knowing that their worker will be randomly drawn from the subset of workers comprising the relevant, displayed distribution. As shown in Appendix Table A.1, these displayed distributions, on average, largely favor male-odd-month workers. This is not surprising given the true underlying average performance gap. Out of the 24 distributions that are displayed in the math version or in the sports version, there are only 3 sets of distributions that favor the female-even-month workers and 1 set of distributions that neither favors the male-odd-month workers nor the female-even-month workers in terms of average performance difference within the subset.

Here, we focus on the decisions over the 20 distributions for which the average performance gap favors the male-odd-month workers.[25] The corresponding decisions can be summarized as follows (and see Appendix Figure A.3 and Appendix Table A.12 for reference). First, employers are less likely to hire from the lower-performing distribution: female workers in the Gender treatment are hired 33% of the time when the average performance gap of the displayed distributions favors male workers. Second, this lower hiring rate is not specific to gender: even-month workers in the Birth Month treatment are hired 29% of the time, significantly lower than the 33% of the time that female workers are hired in the Gender treatment. This suggests that our main result replicates in this environment: there is less hiring of women than men, but it is not driven by gender-specific animus. Third, as in our riskless decisions over workers with identical performances, evidence for in-group preferences is substantial and significant. Female employers are 16 percentage points more likely to hire female workers in the Gender treatment than male

[24] If we breakdown the data by domain, we document significant in-group preferences in both math and sports, but again, only in the riskless decisions. Within each domain, the introduction of risk completely eliminates in-group preferences. See Appendix Table A.11.

[25] As one would expect, the results are much noisier when only considering the 3 distributions for which the average performance gap favors the female-even-month worker (see Appendix Figure A.4 and Appendix Table A.13) and the 1 distribution for which there is no average performance gap (see Appendix Figure A.5 and Appendix Table A.14).

employers are, and even-month employers are 6 percentage points more likely to hire even-month workers in the Birth Month treatment than odd-month employers are. Indeed, the slight preference for hiring female, as compared to even-month, workers appears entirely driven by stronger in-group preferences within the Gender treatment for these Information Stage decisions.

The second set of additional hiring decisions involves employers in the Hiring Stage choosing between pairs of workers in which the female-even-month worker has a better easy round performance than the male-odd-month worker. The corresponding decisions can be summarized as follows (and see Appendix Figure A.6 as well as Appendix Table A.15 for reference). First, female-even-month workers are hired nearly 90% of the time in both of our treatments when the decision does not involve risk, suggesting that any belief-based discrimination against female-even-month workers is not so strong as to induce employers not to hire a worker with a strictly better easy round performance. Second, we again do not observe any evidence consistent with taste-based animus against female workers: female workers in the Gender treatment are not less likely to be hired than even-month workers in the Birth Month treatment. Third, while rates of hiring the female-even-month worker fall with the introduction of risk (just as in the case of equal performances), the female-even-month worker is still hired no less often in the Gender treatment than in the Birth Month treatment even under the risky decisions. And, finally, we do not observe any significant differences according to the demographics of the employer, suggesting that in-group preferences are not so strong as to overcome a desire to hire the worker with the better easy round performance.

3.6 Do our results extend to a simplified, between-subjects environment?

We also document the robustness of our findings in a second, simplified experiment. In this simplified experiment, employers make just a single hiring decision in the Hiring Stage, and we introduce risk as an "excuse" not to hire across-employer, rather than within-employer. This simplified experiment thus allows us to show that our results are not simply driven by bracketing concerns or cognitive load concerns that may arise when employers make many hiring decisions, as in our original experiment.

The design is as follows. First, employers complete the Elicitation of Prior Beliefs Stage. Second, they proceed to a very simplified version of the Information Stage in which they are shown only the final screen of the Information Stage from the original experiment (i.e., the complete distributions of performances of all female-even-month workers and male-odd-month workers) and do not make any decisions about whom to hire from these distributions. Third, employers complete the Elicitation of Posterior Beliefs Stage. Fourth, and most importantly, the Hiring Stage is shortened so that employers only make a single hiring decision. They must choose to hire either a female-even-month worker or a male-odd-month worker, each of whom earned a score of 4 on the easy quiz.

Just as in our original experiment, we randomly assign employers to either the Gender treatment or the Birth Month treatment. To accommodate the one-hiring decision nature of the simplified experiment, we also randomly assign employers to either make this decision in the risky or riskless paradigm. All employers in the riskless paradigm make a hiring decision in which the payoff rule for hiring either worker is the same, while all employers in the risky paradigm make a hiring decision in which the payoff rule is altered so that there is only a 99% chance of receiving the payoff if the female-even-month worker is hired. Thus, in this simplified experiment, the introduction of risk is an across-subject treatment variation, rather than within-subject. Finally, all employers are assigned to the math version of the experiment, reducing the number of employers we need in order to be adequately powered to identify differences across the original and simplified versions of the experiment.

We ran this study on Amazon Mechanical Turk in November 2019 with 1995 employers; it was advertised as a 10-minute academic study with a completion payment of $1.50, plus additional payment based upon the hiring decision made.[26] With the exception of the details described above, the rest of the experimental design exactly replicates the original hiring experiment, including our use of the same available pool of workers.

To compare the decisions made by employers in our simplified experiment to the decisions made by employers in our original experiment, we restrict our analysis to the common hiring decisions across these two experiments: those involving two workers, both with a performance of 4 on the easy math quiz.

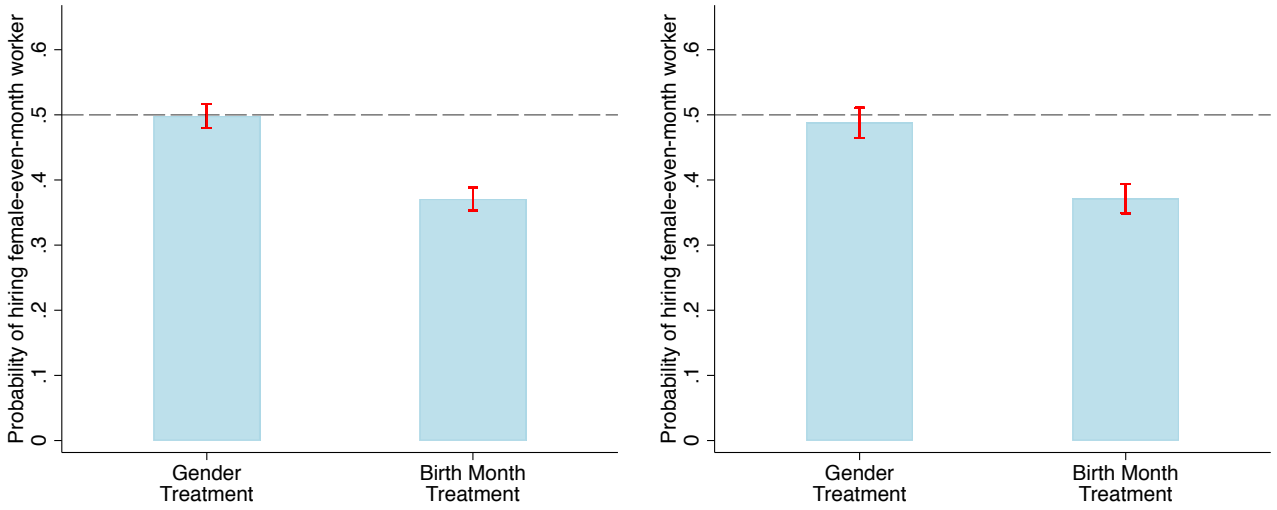
Focusing first on the riskless hiring decisions involving two workers both with a performance of 4 on the easy math quiz, Figure 7 shows that the results are remarkably consistent across the simplified experiment and the original experiment. Most importantly, our main result strongly replicates: women are no less likely to be hired under the Gender treatment than under the Birth Month treatment. In fact, while there is significant discrimination against the female-even-month worker in the Birth Month treatment, women are hired at a rate indistinguishable from 50% within the Gender treatment in both the simplified and original experiments for this particular decision.

We also find a high degree of consistency between the original and simplified experiments when we consider the risky decisions. Appendix Table A.16 replicates Table 1, focusing only on the decisions of interest (riskless and risky hiring decisions involving two workers both with a performance of 4 on the easy math quiz). We present the results for both the simplified experiment (Panel A) and the original experiment (Panel B). The main result is quite similar across both: in both riskless and risky decisions, the Gender treatment has a significant, positive impact on the probability of hiring women.

[26] We recruited 2000 Mturkers, but 5 employers either did not submit the proper evidence of HIT completion or had participated in the original version of this experiment despite our efforts to exclude them. We drop those 5 Mturkers from our sample.

Finally, in considering heterogenous effects, results from our simplified experiment further cast doubt on the robustness of in-group preferences. While in-group preferences were observed in our riskless but not risky decisions in our original experiment, Appendix Figure A.7 and Appendix Table A.17 show that in-group preferences are not even observed within the riskless decisions of our simplified experiment. There are no significant in-group preferences conditional on stated beliefs in the simplified experiment, either in the riskless or the risky decisions.

Figure 7: Comparing the simplified versus original math version: Hiring decisions between workers with same performance levels of 4 and same payoff rule



(a) Simplified Math Version

(b) Original Math Version

Thus, our conclusions drawn from across the many decisions of the original experiment and from the new simplified experiment are remarkably consistent. In our experiments, discrimination against female workers appears to be driven by beliefs, as quite consistently, women are not significantly less likely to be hired when labeled as women than when labeled as even-month workers. This is true even when image concerns are largely reduced, through the introduction of a veil on employers' intentions.

4 Conclusion

We leverage a controlled environment in order to disentangle the motivations behind discrimination against women in a male-typed employment context. We find ample evidence of discrimination against women, as employers, on average, are significantly less likely to hire a woman compared to an equally able man.

This discrimination, however, is not specific to gender. When workers are identified as members of one of two groups, we observe similar levels of discrimination against the lower-performing group regardless of whether they are identified according to gender or birth month. This evidence

points to an important role for beliefs about performance in explaining discriminatory practices. In this way, theories of stereotyping and belief formation may be particularly important for understanding and reducing discrimination.

In our setting, beliefs and tastes work in opposite directions in driving discrimination. When considering the overall results in our context, beliefs push employers to hire women less often, while tastes, if anything, seem to push employers to hire women more often. This is particularly true in the arguably more career relevant domain of math as compared to sports. Future work should investigate whether this is reflective of belief-based discrimination (perhaps particularly in comparison to animus-driven discrimination) as being viewed as appropriate, acceptable, or commonplace.

We also provide evidence on in-group preferences, finding that male employers, more so than female employers, discriminate against female workers in the Gender treatment. Again however, this finding is not specific to gender. Odd-month employers, more so than even-month employers, discriminate against even-month workers. This evidence highlights how discrimination can be mitigated by in-group preferences or activated by out-group animus. We caution, however, that when employers are provided with a veil on their intentions, the addition of risk to the decision, we no longer observe a significant role for in-group preferences. The evidence for in-group preferences is quite mixed across the many decisions we consider.

The value in our paper is showing both how and why it is important to construct careful control comparisons to narrow in on the drivers of discrimination. More specifically, there are three main takeaways from our approach that could help to inform future investigations on the drivers of gender discrimination. First, providing ample information on the distribution of performances for a group — but varying whether the gender of that group is known — is a powerful way to examine the role of beliefs related to average group differences. Average group differences are central to many theories of discrimination; collecting data that speaks directly to their importance provides clearer evidence on these theories. Second, labeling a control group by an arbitrary characteristic — rather than simply removing a gender label — allows one to narrow in on gender-specific channels. It prevents one from attributing taste-based considerations that are driven by in-group preferences to gender. While using birth month to label groups may be an unusual approach, it serves our purposes well. Future work could follow a similar approach with other, more common group associations such as those based upon school affiliation, section or classroom assignment, workplace group, office location, or teams. Third, introducing a veil for intentions by building off of the approach in Exley (2015) — given that decisions outside of the laboratory frequently involve such noise or plausible veils — allows one to speak to the robustness of findings even in rather abstract environments (e.g., by minimizing the role of image concerns, including those related to social desirability or experimenter demand concerns, in driving results).

It is worth emphasizing that we find evidence for discrimination against women even in a

stark environment. In our riskless decisions, employers make a side-by-side decision between a male worker and a female worker with identical resumes, there is little ambiguity to use to justify discrimination, and it is quite cheap to not appear sexist. And, yet, we find that employers, on average, are willing to discriminate against women. In this way, our findings underscore the challenges faced by members of a group that is believed to be lower-performing on average. Mere membership in a lower-performing group — even when this membership is outside of the control of the worker and based on an arbitrary characteristic — is sufficient for discrimination to follow. This challenge, moreover, may be exacerbated for members of a group that is not well-represented among decision-makers. For instance, had all employers been males in our study, the overall level of discrimination against female workers would have been greater. This suggests a need for future work to better understand whether and how shared social identities may reduce discriminatory behavior.

More generally, finding effective remedies for discrimination remains an important challenge for social scientists, policy-makers, and organizational leaders. While future work is clearly needed on this issue, we offer a few suggestions here. Our results suggest that beliefs about average group differences are an important driver of discriminatory behavior. Thus, designing processes that reduce reliance on beliefs about candidates that are informed largely by group membership, and not individual characteristics, should help to limit the extent to which belief-based discrimination occurs in hiring and other workplace decision-making. This could mean collecting more, and more objective, information from each candidate. Acknowledgement of and reflection on beliefs may also be helpful: ask, what beliefs do I hold about the capabilities of this candidate, what informs those beliefs, and would I hold those same beliefs if this candidate were of another gender (race, socioeconomic status, etc.)?

Our work shows the important role for beliefs in a stylized hiring experiment. The advantage of the controlled paradigm is our ability to isolate beliefs as a mechanism. But, understanding how our results would generalize to other contexts is an important question for future research. For instance, if we move to an environment or paradigm in which an employer is likely to believe that women outperform men on average, our results and framework would predict discrimination against men, again driven by beliefs; is this what we would observe empirically? Another important step is moving beyond this type of stylized paradigm. Future work should consider the role for beliefs in contributing to discrimination in other more complex settings, particularly in the field.

References

Altonji, Joseph G., and Rebecca M. Blank. 1999. "Chapter 48 Race and gender in the labor market." In . Vol. 3 of *Handbook of Labor Economics*, 3143 – 3259. Elsevier.

Arrow, Kenneth. 1973. "Discrimination in Labor Markets." , ed. Orley Ashenfelter and Albert Rees, Chapter The Theory of Discrimination. Princeton, NJ:Princeton University Press.

Ayres, Ian, and Peter Siegelman. 1995. "Race and gender discrimination in bargaining for a new car." *The American Economic Review*, 304–321.

Azmat, Ghazala, and Barbara Petrongolo. 2014. "Gender and the labor market: What have we learned from field and lab experiments?" *Labour Economics*, 30: 32–40.

Baert, Stijn, Ann-Sophie De Pauw, and Nick Deschacht. 2016. "Do employer preferences contribute to sticky floors?" *ILR Review*, 69(3): 714–736.

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

Bartoš, Vojtěch, Michal Bauer, Julie Chytilová, and Filip Matějka. 2016. "Attention Discrimination: Theory and Field Experiments with Monitoring Information Acquisition." *American Economic Review*, 106(6): 1437–1475.

Becker, Gary S. 1957. *The Economics of Discrimination.* Chicago, IL:University of Chicago Press.

Bertrand, Marianne. 2011. "New perspectives on Gender." *Handbook of Labor Economics*, 4: 1543–1590.

Bertrand, Marianne, and Esther Duflo. 2017. "Field experiments on discrimination." *Handbook of Economic Field Experiments*, 1: 309–393.

Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination." *The American Economic Review*, 94(4): 991–1013.

Black, Sandra E, and Philip E Strahan. 2001. "The division of spoils: rent-sharing and discrimination in a regulated industry." *The American Economic Review*, 91(4): 814–831.

Blau, Francine D, and Lawrence M Kahn. 2017. "The gender wage gap: Extent, trends, and explanations." *Journal of Economic Literature*, 3: 789–865.

Bohnet, Iris, Alexandra van Geen, and Max Bazerman. 2016. "When Performance Trumps Gender Bias: Joint Versus Separate Evaluation." *Management Science*, 62(5): 1225–1234.

Bohren, J Aislinn, Alex Imas, and Michael Rosenberg. 2017. "The Dynamics of Discrimination: Theory and Evidence." *Working Paper*.

Bordalo, Pedro, Katherine B. Coffman, Nicola Gennaioli, and Andrei Shleifer. 2019. "Beliefs about Gender." *American Economic Review*, 109(3): 739–773.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. 2016. "Stereotypes." *The Quarterly Journal of Economics*, 131(4): 1753–1794.

Bowles, Hannah Riley, Linda Babcock, and Lei Lai. 2007. "Social incentives for gender differences in the propensity to initiate negotiations: Sometimes it does hurt to ask." *Organizational Behavior and Human Decision Processes*, 103(1): 84–103.

Brewer, Marilynn B. 1979. "In-group bias in the minimal intergroup situation: A cognitive-motivational analysis." *Psychological bulletin*, 86(2).

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

Card, David, Ana Rute Cardoso, and Patrick Kline. 2016. "Bargaining, sorting, and the gender wage gap: Quantifying the impact of firms on the relative pay of women." *Quarterly Journal of Economics*, 131(2): 633–686.

Castillo, Marco, and Ragan Petrie. 2010. "Discrimination in the lab: Does information trump appearance?" *Games and Economic Behavior*, 68.

Castillo, Marco, Ragan Petrie, Maximo Torero, and Lise Vesterlund. 2013. "Gender differences in bargaining outcomes: A field experiment on discrimination." *Journal of Public Economics*, 99: 35–48.

Charles, Kerwin Kofi, and Jonathan Guryan. 2008. "Prejudice and Wages: An Empirical Assessment of Becker's The Economics of Discrimination." *Journal of Political Economy*, 116(5): 773–809.

Charles, Kerwin Kofi, and Jonathan Guryan. 2011. "Studying discrimination: Fundamental challenges and recent progress." *Annual Review of Economics*, 3(1): 479–511.

Chen, Roy, and Yan Chen. 2011. "The potential of social identity for equilibrium selection." *The American Economic Review*, 101(6): 2562–2589.

Chen, Yan, and Sherry Xin Li. 2009. "Group identity and social preferences." *The American Economic Review*, 99(1): 431–457.

Chen, Yan, Sherry Xin Li, Tracy Xiao Liu, and Margaret Shih. 2014. "Which hat to wear? Impact of natural identities on coordination and cooperation." *Games and Economic Behavior*, 84: 58–86.

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

Croson, Rachel, and Uri Gneezy. 2009. "Gender Differences in Preferences." *Journal of Economic Literature*, 47(2): 448–474.

Dana, Jason, Roberto A. Weber, and Jason Xi Kuang. 2007. "Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness." *Economic Theory*, 33: 67–80.

Danilov, Anastasia, and Silvia Saccardo. 2017. "Disguised Discrimination." *Working Paper*.

De Paola, Maria, and Vincenzo Scoppa. 2015. "Gender discrimination and evaluators' gender: evidence from Italian academia." *Economica*, 82(325).

Dickinson, David L., and Ronald L. Oaxaca. 2009. "Statistical Discrimination in Labor Markets: An Experimental Analysis." *Southern Economic Journal*, 76(1): 16–31.

Dickinson, David L, and Ronald L. Oaxaca. 2014. "Wages, Employment, and Statistical Discrimination Evidence from the laboratory." *Economic Inquiry*, 52(4): 1380–1391.

Doleac, Jennifer L., and Luke C.D. Stein. 2013. "The Visible Hand: Race and Online Market Outcomes." *The Economic Journal*, 123.

Exley, Christine L. 2015. "Excusing Selfishness in Charitable Giving: The Role of Risk." *Review of Economic Studies*, 83(2): 587–628.

Fershtman, Chaim, and Uri Gneezy. 2001. "Discrimination in a segmented society: An experimental approach." *The Quarterly Journal of Economics*, 116(1): 351–377.

Gino, Francesca, Michael I. Norton, and Roberto A. Weber. 2016. "Motivated Bayesians: Feeling Moral While Acting Egoistically." *Journal of Economic Perspectives*, 30(3): 189–212.

Gneezy, Uri John List, Michael K. Price. 2012. "Toward an Understanding of Why People Discriminate: Evidence from a Series of Natural Field Experiments." *NBER Working Paper*.

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

Goldin, Claudia, and Cecilia Rouse. 2000. "Orchestrating Impartiality: The Impact of" Blind" Auditions on Female Musicians." *The American Economic Review*, 90(4): 715–741.

Guryan, Jonathan, and Kerwin Kofi Charles. 2013. "Taste-based or Statistical Discrimination: The Economics of Discrimination Returns to its Roots." *The Economic Journal*, 123(572).

Haisley, Emily C., and Roberto A. Weber. 2010. "Self-serving interpretations of ambiguity in other-regarding behavior." *Games and Economic Behavior*, 68: 614–625.

Halberda, Justin, Ryan Ly, Jeremy B Wilmer, Daniel Q Naiman, and Laura Germine. 2012. "Number sense across the lifespan as revealed by a massive Internet-based sample." *Proceedings of the National Academy of Sciences*, 109(28): 11116–11120.

Krupka, Erin L., and Roberto A. Weber. 2013. "Identifying social norms using coordination games: Why does dictator game sharing vary?" *Journal of the European Economic Association*, 11(3): 495–524.

Kunda, Ziva. 1990. "The Case for Motivated Reasoning." *Psychological Bulletin*, 108(3): 480–498.

Lei, Lai, and Linda C. Babcock. 2013. "Asian Americans and workplace discrimination: The interplay between sex of evaluators and the perception of social skills." *Journal of Organizational Behavior*, 34(3): 310–326.

List, John A. 2004. "The nature and extent of discrimination in the marketplace: Evidence from the field." *The Quarterly Journal of Economics*, 119(1): 48–89.

Milkman, Katherine L., Modupe Akinola, and Dolly Chugh. 2012. "Temporal Distance and Discrimination: An Audit Study in Academia." *Psychological Science*, 23(7): 710–717.

Milkman, Katherine L, Modupe Akinola, and Dolly Chugh. 2015. "What happens before? A field experiment exploring how pay and representation differentially shape bias on the pathway into organizations." *Journal of Applied Psychology*, 100(6).

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

Moss-Racusin, Corinne A, John F Dovidio, Victoria L Brescoll, Mark J Graham, and Jo Handelsman. 2012. "Science faculty's subtle gender biases favor male students." *Proceedings of the National Academy of Sciences*, 109(41): 16474–16479.

Neumark, David, Roy J. Bank, and Kyle D. Van Nort. 1996. "Sex Discrimination in Restaurant Hiring: An Audit Study." *Quarterly Journal of Economics*, 111(3): 915–941.

Niederle, Muriel. 2016. "Gender." In *Handbook of Experimental Economics*. Vol. 2, , ed. John Kagel and Alvin E. Roth, 481–553. Princeton University Press.

Olivetti, Claudia, and Barbara Petrongolo. 2016. "The Evolution of Gender Gaps in Industrialized Countries." *Annual Review of Economics*, 8(1): 405–434.

Paolacci, Gabriele, Jesse Chandler, and Panagiotis G. Ipeirotis. 2010. "Running experiments on amazon mechanical turk." *Judgment and Decision Making*, 5(5): 411–419.

Phelps, Edmund S. 1972. "The statistical theory of racism and sexism." *The American Economic Review*, 62(4): 659–661.

Rand, David G. 2012. "The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments." *Journal of theoretical biology*, 299: 172–179.

Rao, Gautam. 2018. "Familiarity Does Not Breed Contempt: Generosity, Discrimination and Diversity in Delhi." *Working Paper*.

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

Riach, P. A., and J. Rich. 2002. "Field Experiments of Discrimination in the Market Place." *The Economic Journal*, 112(483).

Sarsons, Heather. 2017*a*. "Interpreting Signals in the Labor Market: Evidence from Medical Referrals." *Working Paper*.

Sarsons, Heather. 2017*b*. "Recognition for Group Work: Gender Differences in Academia†." *American Economic Review: Papers & Proceedings*, 107(5): 141–145.

Tajfel, Henri, Michael G Billig, Robert P Bundy, and Claude Flament. 1971. "Social categorization and intergroup behaviour." *European Journal of Social Psychology*, 1(2): 149–178.