BERLIN SCHOOL OF ECONOMICS
DISCUSSION PAPERS

Discussion Paper #35

March 2024

Explicit and Implicit Belief-Based Gender Discrimination: A Hiring Experiment

Kai Barron
Ruth Ditlmann
Stefan Gehrig
Sebastian Schweighofer-Kodritsch

Explicit and Implicit Belief-Based Gender Discrimination: A Hiring Experiment*

Kai Barron
WZB Berlin

Ruth Ditlmann
Hertie School Berlin

Stefan Gehrig
WZB Berlin

Sebastian Schweighofer-Kodritsch
Humboldt-Universität zu Berlin

November 14, 2023

Abstract

This paper studies a key element of discrimination, namely when stereotypes translate into discriminatory actions. Using a hiring experiment, we rule out taste-based discrimination by design and test for the presence of two types of belief-based gender discrimination. We document evidence of *explicit discriminators*—individuals who are willing to discriminate even when their hiring choices are highly revealing of their gender-biased beliefs. Crucially, we also identify *implicit discriminators*—individuals who do not discriminate against women when taking a discriminatory action is highly revealing of their biased beliefs, but do discriminate against women when their biased motive is obscured. Our analysis highlights the central role played by features of the choice environment in determining whether and how discrimination will manifest. We conclude by discussing the implications for policy design.

JEL Codes: D90, J71, D83

Keywords: Discrimination, Hiring Decisions, Gender, Beliefs, Experiment, Aversive Sexism.

*The authors would like to thank Martin Abel, Vojtech Bartos, Katherine Coffman, Alex Coutts, Tom Cunningham, Jon de Quidt, Tilman Fries, Thomas Graeber, Simone Häckl, Kareem Haggag, Lea Heursen, Alex Imas, Dorothea Kübler, Yves Le Yaouanq, Heather Sarsons, Julia Schmieder, Florian Schneider, Alice Solda, Martin Spann, Robert Stüber, Müge Süer, Roel van Veldhuizen and the audiences at the WZB Brown Bag Seminar, the ESA Global 2020 Meeting, the CRC TRR 190 Ohlstadt 2020 Retreat, the 2021 European Winter Meeting of the Econometric Society, the Bergen-Berlin Behavioral Economics Spring Workshop 2022, the MiddExLab Seminar, Stockholm University's SOFI Labour Economics Seminar, UC Davis's Behavioral Economics Seminar, and the 2nd Berlin Workshop on Empirical Public Economics: Gender Economics for helpful comments. We thank the WZB for generously funding this project through its interdisciplinary "seed money" programme. Barron and Schweighofer-Kodritsch gratefully acknowledge financial support by the *Deutsche Forschungsgemeinschaft* through CRC TRR 190 (project number 280092119). The data and software code to reproduce all analyses are available at `https://github.com/stefgehrig/implicitgender`.

1 Introduction

Discrimination is a critically important policy issue around the world.[1] When one individual receives preferential treatment over another on the basis of their appearance or group identity, this violates basic meritocratic principles. In the labor market, it can also be inefficient, resulting in a less productive workforce, and tends to reinforce inequality within society. In relation to gender, which is the focus of this paper, a substantial body of work has provided evidence that discrimination plays an important role in generating the gender gap observed in labor market outcomes.[2] It is, therefore, crucial for the design of effective policies to be able to understand the drivers of this discrimination.

Traditionally, the economics literature has distinguished between *taste-based* discrimination (Becker, 1957) and *statistical* discrimination (Phelps, 1972; Arrow, 1973). When engaging in taste-based discrimination, an employer prefers to avoid interacting with or being associated with individuals from a particular group. Statistical discrimination involves an employer with imperfect information treating individuals differently due to (accurate) beliefs about *statistical* differences in ability, skill, or productivity between groups. This dichotomy looks at discrimination through the lens of the subjective expected utility framework, distinguishing between discrimination based on preferences and discrimination based on accurate beliefs.

Recent work suggests that this taxonomy may be too narrow in important ways. First, it is common for discrimination to emanate from *inaccurate beliefs* due to stereotypes or biases in belief formation (see, e.g., Judd and Park, 1993; Hilton and Von Hippel, 1996; Heilman, 2012; Bordalo, Coffman, Gennaioli, and Shleifer, 2016; Bohren et al., 2022; Mengel and Campos-Mercade, 2023). Second, research on "aversive racism" in social psychology argues that individuals may face a tension between their discriminatory stereotypes/preferences and their desire to comply with a social norm against discrimination—e.g., to maintain a positive social or self-image (e.g., Snyder, Kleck, Strenta, and Mentzer, 1979; Banaji and Greenwald, 1995; Hodson, Dovidio, and Gaertner, 2010). According to this "aversive racism" framework, which we discuss in more detail in Appendix B, such individuals are often unaware of their own bias. They actively try to avoid appearing biased and tend to only discriminate in situations where they can find a plausible alternative justification for their decisions. One proposed mechanism for these justifications is that evaluators shift their decision-making criteria ex-post in order to mask their biased decisions (see, e.g., Norton, Vandello, and Darley, 2004; Uhlmann

[1] For a review of the economics literature on discrimination, see: Riach and Rich (2002), Charles and Guryan (2011), Lane (2016), Bertrand and Duflo (2017), Blau and Kahn (2017) and Neumark (2018).

[2] Evidence has been documented in a diverse range of contexts including *bargaining* (Ayres and Siegelman, 1995; Bowles, Babcock, and Lai, 2007; Small, Gelfand, Babcock, and Gettman, 2007), *hiring* (Jowell and Prescott-Clarke, 1970; Newman, 1978; McIntyre, Moberg, and Posner, 1980; Yinger, 1986; Riach and Rich, 1987; Glick, Zion, and Nelson, 1988; Neumark, Bank, and Van Nort, 1996; Biernat and Kobrynowicz, 1997; Goldin and Rouse, 2000; Bertrand and Mullainathan, 2004; Reuben, Sapienza, and Zingales, 2014; Bohnet, Van Geen, and Bazerman, 2015; Milkman, Akinola, and Chugh, 2015; Kübler, Schmid, and Stüber, 2018; Bohren, Haggag, Imas, and Pope, 2022; Coffman, Exley, and Niederle, 2021), *referrals, promotions, and recognition for (group-)work* (Isaksson, 2018; Coffman, Flikkema, and Shurchkov, 2021; Sarsons, 2019; Hengel, 2022; Card, DellaVigna, Funk, and Iriberri, 2020; Sarsons, Gërxhani, Reuben, and Schram, 2021).

and Cohen, 2005). Recent work in economics has started to build on these ideas, exploring the implications of discriminatory preferences/stereotypes that individuals themselves may be unaware of (see, e.g., Bertrand, Chugh, and Mullainathan, 2005; Bertrand and Duflo, 2017; Carlana, 2019; Alesina, Carlana, La Ferrara, and Pinotti, 2018) and exploring how discriminatory preferences/stereotypes may be more likely to influence behavior in contexts where the discrimination is less obvious (see, e.g., Bohnet et al., 2015). This literature typically uses one of two approaches. The first thread of research uses the implicit association test (IAT) to measure unconscious bias and then examines the implications of this bias. The second compares the aggregate level of bias between scenarios where discrimination is more or less obvious—for example, Bohnet et al. (2015) examine the gender bias when job candidates are evaluated jointly (more obvious) or separately (less obvious). In an important contribution to this literature, Cunningham and de Quidt (2023) propose a comprehensive theoretical framework that organizes and clarifies these overlapping ideas around implicit preferences. Additionally, they provide a description of how to identify implicit preferences from choice behavior.

We contribute to this literature by providing the first empirical study that draws on the Cunningham and de Quidt (2023) theoretical framework to identify both explicit and implicit gender discrimination in the context of labor market choices. To provide a sense of what we mean by explicit and implicit discrimination, consider an employer who will always choose to hire a woman over a man when they hold exactly the same qualifications, but will always choose to hire a man over a woman when they hold different qualifications. Such an employer discriminates explicitly against men, but implicitly against women. This is because, in the first scenario, where candidates are *equally qualified*, discrimination is obvious, revealing an explicit preference. In the second scenario, candidates are *differently qualified*, and their ranking is more subjective, making discrimination less obvious and revealing an implicit preference.[3] This logic captures the basic intuition of the two types of discrimination we study in this paper.

To detect implicit discrimination, we need to observe multiple hiring decisions by the same individual, where the job candidates' attributes vary systematically in a particular way across hiring decisions. This is difficult to achieve with naturally occurring data. We therefore design a hiring experiment. This also allows us to construct a choice setting that completely rules out classical taste-based discrimination. This means that we are able to focus on isolating different forms of belief-based discrimination. In this, we join the rapidly growing contemporary literature that studies the central role of (possibly inaccurate) beliefs or stereotypes in generating

[3]The core assumption of the Cunningham and de Quidt (2023) framework, which they refer to as "dilution", is that the influence of an implicit preference for an attribute (here, gender) increases when it is "mixed" with other attributes (here, qualifications). When a hiring choice only involves a difference in gender (i.e., equally qualified candidates), the choice is easily attributed to preferences/beliefs about gender. When the hiring choice involves candidates that differ on many attributes, various explanations or justifications are possible for the hiring choice, and discrimination on gender is less obvious.

different behavior towards men and women.[4]

Our experiment simulates the main features of a hiring scenario, but is carefully constructed to allow us to identify both explicit and implicit belief-based discrimination. To obtain "job candidates" for the hiring decisions in the main experiment, we collected performance information from a first group of 80 participants who completed a series of quizzes. In the main hiring experiment, 240 participants take the role of "employers" and make a series of hiring decisions between pairs of job candidates.

When making a hiring decision, the employer observes a "mini-CV" for each of the two candidates. The mini-CV contains information about the candidate's gender and two possible qualifications. A qualification takes the form of a "certificate" that is awarded to candidates who score in the top 30% in a particular qualification task. There are two qualification tasks – a general knowledge task and a word search task. These qualification tasks are distinct from the job task, which is a logic task. Employers are incentivized to hire the better job candidate – they receive a fixed payment if they hire the candidate that performed better in the job task. To allow employers to indicate indifference between the two candidates, we introduce the following design feature. After making their initial choice in a hiring decision, an employer can "sell" their choice for a small payment of 0.10€. If they do this, one of the two candidates is selected at random.

Our analysis consists of two main parts. First, we identify gender bias at the aggregate level in different types of hiring decisions. This analysis looks at whether men are hired more often than women on average, holding qualifications constant. These aggregate bias exercises are similar to what is done in much of the extant literature. However, our study differs from the majority of this previous work in two respects: (i) we examine this bias for three distinct classes of hiring decisions, namely those in which candidates are *equally qualified*, *differently qualified*, and those in which one candidate is *more qualified*, and (ii) we rule out taste-based discrimination, which implies that the bias is due to (inaccurate) beliefs. Second, the main focus and contribution of our paper is that we go beyond these aggregate bias exercises and exploit the within-participant data that we collect to detect implicit discrimination. The aggregate bias exercises can allude to implicit discrimination (by comparing the magnitude of bias in scenarios where discrimination is obvious vs not obvious), but cannot identify it cleanly—this requires observing multiple within-individual decisions across different hiring scenarios.

To identify aggregate level gender bias, we compare pairs of hiring decisions made by employers between a male and a female candidate. The "qualification profiles" of the two

[4]This area of research, focused on the role of (biased) beliefs as a driver of discrimination, has seen extremely rapid growth in the last few years (see, e.g., Bordalo et al., 2016; Bohren, Imas, and Rosenberg, 2019; Coffman et al., 2021; Bordalo, Coffman, Gennaioli, and Shleifer, 2019; Bohren et al., 2022; Mengel and Campos-Mercade, 2023; Coffman, Collis, and Kulkarni, 2021; Coffman, Araya, and Zafar, 2021; Erkal, Gangadharan, and Koh, 2021; Lepage, 2021a; Bohren, Hull, and Imas, 2022; Esponda, Oprea, and Yuksel, 2022). Aside from this recent wave of papers, Bohren et al. (2022) note that very little empirical work in economics between 1990 and 2018 considered the role of inaccurate beliefs in discrimination. Their review of the literature indicates that only five of the 105 papers they identified in top economics journals tested for inaccurate beliefs. Collectively, the recent literature suggests this is an important omission.

candidates are held constant across the pair of decisions, but the gender of the two candidates is reversed. Therefore, in one decision, the male candidate has qualification profile A and the female candidate has qualification profile B, and in the other, this is reversed. In our experiment, these profiles can either indicate that a candidate has the certificate for the knowledge task, the certificate for the word task, or not indicate any certificate. Aggregating the decisions made in these pairs of hiring decisions enables us to cleanly attribute any significant difference in how often the male vs. female candidate is hired to the gender of the candidates and implies gender discrimination. One major advantage of our approach is that we are able to identify discrimination purely behaviorally, without imposing any assumptions on employers' subjective beliefs. This is because any difference in hiring men vs. women must be due to gender, since all other information is held constant and balanced across the pair of decisions.

We find the following patterns of belief-based gender bias at the aggregate level. First, whenever both candidates are *equally qualified*, there is significant discrimination against women. Specifically, employers hire the male candidate in 57% of these hiring decisions, while the female candidate is hired in only 43% of them, despite having identical qualifications. Since the only difference between the two candidates is their gender, this indicates that some employers view gender as providing an informative signal about performance and are willing to act on this even when it is obvious (explicit) that they are using gender information in this way. Since such decisions cannot be attributed to any other candidate characteristics, we term this "explicit" discrimination. Here, it is "statistically inaccurate" (we borrow this terminology from Bohren et al., 2022), since female candidates were not objectively out-performed by male candidates; in this sense, our experiment rules out (accurate) statistical discrimination.

Second, whenever the two candidates are *differently qualified* (i.e., each candidate has a different certificate), there is again a significant gender bias and hence discrimination against women. Interestingly, the magnitude of the gender bias against women here is approximately as large as in the hiring decisions where candidates are *equally qualified*, with the male candidate hired in approximately 58% of such hiring decisions. At surface level, the similarity in the magnitudes of the aggregate bias in these two types of hiring decisions appears to suggest that implicit discrimination may not be a substantial factor in our sample. Importantly, however, these aggregate patterns reflect the net effect of pooling all forms of discrimination in a particular hiring context (e.g., discrimination in favor of and against women). Therefore, as we discuss below, they are, in fact, consistent with either an extremely high or low prevalence of implicit bias. This is why we take advantage of the rich within-participant information we collect to detect implicit discrimination.

Third, whenever one candidate is *more qualified* than the other (i.e., one has a certificate, the other has none), the more qualified candidate is always hired at approximately the same rate of 75 to 85%, irrespective of gender; i.e., there is no significant gender bias. This demonstrates that, at the aggregate level, participants in our sample do not always display a gender bias—when the comparison of qualifications between the two candidates contains a strong signal, the gender bias dissipates.

This analysis of aggregate bias does not take full advantage of the fact that each employer makes multiple hiring decisions. To go beyond aggregate bias and achieve the central objective of our paper, namely detecting implicit discrimination, we have every employer in the experiment make a series of nine carefully constructed hiring decisions in which we systematically vary the attributes on the CVs of each of the candidates. We are able to use the pattern of behavior displayed by a given participant across these hiring scenarios to better understand what is driving the patterns of behavior observed at the aggregate level. In particular, we can evaluate whether employers engage in explicit or implicit gender discrimination.

To detect implicit discrimination, we start by recalling from above that an employer who discriminates by gender when candidates are equally qualified engages in explicit gender discrimination. Then, intuitively, an employer engages in implicit gender discrimination if we observe a gender preference reversal between their hiring choices when discrimination is more versus less obvious, i.e., between hiring scenarios where the two candidates are equally versus differently qualified. In this way, our study serves as a proof of concept for the detection of nuanced forms of discrimination that may be pooled together when examining net aggregate bias. To do this, we first classify individual employers into explicit discrimination types based on their hiring choices when candidates are equally qualified. In these two hiring choices, candidates hold exactly the same qualifications and differ only on their gender. Participants are classified as discriminating explicitly against women, against men, or not discriminating explicitly by revealing indifference when candidates differ only in their gender.[5] A final fourth category contains those who display mixed behavior when discrimination is explicit, choosing the male candidate in one hiring choice and the female candidate in the other. We then examine the hiring choices of these different explicit discrimination types when candidates are differently qualified to shed light on their "implicit" gender bias. This identification approach is in line with the formal framework for identifying implicit preferences more generally proposed by Cunningham and de Quidt (2023).

In our explicit discrimination classification exercise, we find that the majority of participants do not display an explicit bias against either gender. However, approximately 40% do, and these employers are twice as likely to discriminate explicitly against women (25.8%) as against men (12.9%). Most strikingly, however, examining the hiring choices of these two types of employers when they choose between candidates that are *differently qualified* reveals a substantial gender bias against women for *both* types. Surprisingly, this gender bias against women is of a very similar magnitude for both those who discriminate explicitly against men and those who discriminate explicitly against women. We, therefore, identify a significant share of employers who discriminate explicitly against men (i.e., in favor of women) in decisions where discrimination is obvious, yet discriminate against women in decisions where

[5]For example, an employer is classified as discriminating explicitly against women here if they choose the male candidate in both of the hiring decisions they face where the candidates (male vs. female) are equally qualified and choose not to sell at least one of them.

discrimination is opaque, i.e., implicitly.[6] Our analysis indicates that 4-8% of all employers discriminate implicitly against women in our experiment.

In Appendix D, we describe the results from several robustness exercises. For example, we conduct a series of placebo exercises that show that our results are unlikely to be generated by noisy choice. We also provide evidence that our results are not driven by a "balancing heuristic," whereby employers try to balance the gender of their hiring portfolio by equalizing the number of men and women they hire across choices. In Section 6.1, we provide an additional test for implicit discrimination by checking for non-transitive cycles in choice corresponding to the "right triangles" of Cunningham and de Quidt (2023). In line with our main analysis, we again find evidence of implicit discrimination against women.

What is driving this implicit discrimination? The pattern of behavior that we observe is consistent with a stereotype that men are better at logic tasks, which could lead employers to form gender-biased beliefs (see also Bordalo et al., 2019). By eliciting the beliefs of the job candidates, we find that these candidates themselves believe that men perform better in the logic task (i.e., the job task), which supports this explanation. The implicit discrimination that we identify, therefore, appears to be due to a combination of: (i) an aversion to displaying overtly discriminatory behavior, and (ii) holding (mistaken) stereotypes. At heart, it involves a tension between holding these gender-based beliefs (stereotypes) and taking actions that might reveal these (stigmatized) beliefs. This reluctance by some participants to discriminate overtly against women, despite holding a biased stereotype, may emanate from the fact that gender discrimination against women is stigmatized in certain segments of the population. Due to image concerns, employers may wish to signal that they do not discriminate, irrespective of any beliefs they may hold about who is the better candidate. While our experiment cannot definitively demonstrate that the implicit discrimination we observe is driven by self or social image concerns, our experimental design contains additional features that allow us to examine whether employer behavior is consistent with this explanation. Specifically, we elicit employer beliefs about the relative importance of the two possible qualifications (word or knowledge) for predicting job performance. The belief data suggests that employers partially rationalize gender-biased choices as being qualification-driven by "redefining merit" ex-post (see, e.g., Uhlmann and Cohen, 2005). This involves adjusting one's beliefs about the relative importance of each of the qualifications for predicting performance in the job task to justify one's gender-based decisions. Intuitively, an employer that chooses to hire a male candidate who has a knowledge certificate over a female candidate with a word certificate might convince themself that the knowledge certificate is more valuable to justify their choice. When instead facing a choice between a male candidate who has a word certificate and a female candidate with a knowledge certificate, the employer might also choose the male candidate, convincing themself that now the word certificate is more valuable. Hence, the perceived value of the qualification

[6] By contrast, employers that engage in "explicit non-discrimination" by revealing indifference in both choices where candidates are equally qualified also do not display a gender bias on average when candidates are differently qualified.

certificate is adjusted ex-post and depends on who holds it.

The paper proceeds as follows. Section 2 discusses the related literature. Section 3 describes the experimental design and Section 4 outlines our main hypotheses. Section 5 presents the main results, while Section 6 details several additional findings. Finally, Section 7 discusses the policy implications of our study and offers conclusions.

2 Related Literature

Our paper lies at the intersection between psychology and economics, incorporating ideas, concepts, and methodological tools from both fields. While it is written using the language of economics, in our view it offers a contribution to the literature in both disciplines.

In economics, our paper relates most closely to the recent literature examining belief-based forms of discrimination (see Reuben et al., 2014; Bordalo et al., 2016; Bohren et al., 2019; Sarsons, 2019; Sarsons et al., 2021; Coffman et al., 2021; Bohren et al., 2022; Eyting, 2022; Rackstraw, 2022; Mengel and Campos-Mercade, 2023; Hübert and Little, 2023; Esponda, Oprea, and Yuksel, 2023). This literature extends the notion of statistical discrimination (Phelps, 1972; Arrow, 1973) and shows that (accurate or inaccurate) group stereotypes can be extremely harmful in generating the differential treatment of near-identical individuals, simply because they belong to different groups.

One strand of this literature explores how group stereotypes can result in discrimination. Here, the stereotypes are taken to be exogenous. While they may be inaccurate, conditional on holding the stereotype, discrimination may be considered "rational". Reuben et al. (2014) show that men are hired at twice the rate of women to perform an arithmetic task in an experimental market, despite the fact that both genders perform equally well on average. The authors attribute this gender gap in hiring to stereotypes, using the implicit association test (IAT) to show that employers' IAT scores were strongly associated with the gender gap in their evaluation of candidates. Similarly, Sarsons et al. (2021) present compelling field and experimental evidence showing that women are given less credit for group work compared to men and that this can be attributed to gender stereotypes. Coffman et al. (2021) show that, on average, employers prefer to hire a male worker over a female worker with an identical resume in a stereotypically-male task. The authors show that similar group-based discrimination is generated when groups are defined according to some arbitrary characteristic—employers simply prefer to hire individuals from the group that (they believe) performs better on average. This highlights the important role played by employers' stereotypical beliefs about groups for the treatment of individuals. In an important contribution to this thread of literature, Bohren et al. (2019) demonstrate that examining dynamic behavior can help to isolate the source of discrimination. Using a field experiment with dynamic evaluations, they show that the observed pattern of discriminatory evaluation behavior is consistent with a belief-based account. Bohren et al. (2022) consider the problem of isolating the source of discrimination in a static

setting, providing evidence that *inaccurate statistical discrimination* is an important source of discrimination that has been largely overlooked by the preceding literature in economics.

A second strand of literature examines how biased beliefs about groups (i.e., stereotypes) or about individual members of groups may arise. This literature examines deviations from rationality in the formation of beliefs. Bordalo et al. (2016) provide a theory of stereotypes based on Kahneman and Tversky's representativeness heuristic in which decision-makers overweight representative types in the group. This can result in belief distortions, such as overemphasizing (perceived) group differences, leading to the polarization of beliefs about group characteristics. Esponda et al. (2022) extend this thread of work by showing that stereotypes about group differences distort the interpretation of new information about individuals from those groups, resulting in biased belief updating and (irrationally) enlarging the discriminatory gap in the evaluation of identical individuals who are members of different groups.[7] Different from this work on cognitive biases emanating from the representativeness heuristic, recent studies explore the relationship between taste-based and belief-based discrimination by considering the role of motivated reasoning. Eyting (2022) provides evidence of discrimination due to motivated reasoning, showing that it is distinct from classical notions of taste-based and statistical discrimination. In related work, Rackstraw (2022) shows that employers update their beliefs more when they are consistent with their pre-existing racial stereotypes.

We contribute to both of these strands of literature in economics, but we also differ from both strands in several important ways. Similar to the first strand, we also examine how the stereotypes individuals hold about groups may translate into discrimination. A key difference is that in our study, we investigate how features of the decision environment can trigger whether a stereotype manifests as discrimination or not. Specifically, we identify belief-based implicit discrimination. An individual who engages in implicit discrimination may always hold stereotypical beliefs, but only act on them in scenarios where their discrimination is not clearly revealed by their actions. This distinguishes our study from the studies in the first thread, which tend not to focus on how features of the decision-making environment influence whether stereotypes translate into discriminatory actions. Perhaps an exception to this is Coffman et al. (2021), who ask whether gender per se is an important group feature when considering the role of gender stereotypes in the hiring discrimination they observe. Our study differs in that the decision-making feature we consider is the degree to which the action reveals discrimination rather than whether the groups are defined by gender or some other characteristic.

While our study is more squarely situated within this first strand, we also contribute to the second strand examining the formation of discriminatory beliefs. Specifically, we examine how individuals might "redefine merit" by adjusting their beliefs about the informativeness

[7]Mengel and Campos-Mercade (2023) illustrate a different, but intuitive, channel for why beliefs about group differences may result in biased beliefs about individual group members that persist despite opportunities to learn about the individual. The authors show that when employers are conservative and therefore do not update their beliefs enough in response to new information, they will discriminate more against an individual from a group that (they believe) performs worse on average. This is because the employer overweights the group information and underweights the individual-specific information.

of the qualifications held by the two candidates that they are evaluating in order to justify their discriminatory hiring decision—they might convince themselves ex-post that whatever qualification the male candidate holds is more informative about job performance and therefore their decision to hire the man had nothing to do with gender. This is a form of motivated reasoning and shows how belief distortion may play a role in our setting. We differ from the contemporary studies in the economics literature examining motivated reasoning in that we consider this very precise form of motivated reasoning, namely adjusting the subjective valuation of qualifications ex-post (i.e., *motivated justifications* for discriminatory decisions).

In social psychology, our study is closely related to the ideas developed in the "aversive racism" framework (see, e.g., Kovel, 1970; Gaertner and Dovidio, 2000). A central idea of this framework is that there is often a discrepancy between the deeply held racial views that individuals hold and their openly stated attitudes and actions—individuals may hold racially biased stereotypes, but wish not to appear racist. The ideas underlying the "aversive racism" framework are conceptually very close to the notion of implicit discrimination considered in this paper and Cunningham and de Quidt (2023). Indeed, we view our paper as building on the intellectual heritage provided by this rich literature in social psychology. In Appendix B, we provide an overview of this literature and discuss how our paper relates to it.[8] In short, we view our paper as making several contributions relative to this body of work. One key contribution is to help introduce the concept of aversive sexism to the economics literature, using the terminology, theories, and methods of experimental economics. This is valuable because re-interpreting these ideas through a completely different disciplinary lens is non-trivial. In doing this, we provide a form of empirical proof-of-concept for how implicit discrimination can be identified by drawing on the behavioral identification approach proposed by Cunningham and de Quidt (2023).

In addition to incorporating existing ideas from psychology into behavioral economics, our study also contributes directly to the social psychology literature in several important ways: First, we use a within-subject design to systematically detect both explicit and implicit gender discrimination (aversive sexism) in hiring decisions. In doing this, we relate closely to Uhlmann and Cohen (2005), who show that evaluators adjust their perceptions of male and female candidates' credentials ex-post to justify hiring them in gender-stereotypical jobs. Similarly, Norton et al. (2004) discuss the role of *casuistry*, showing that male participants justify hiring male candidates by inflating the importance of whatever qualification the male candidate has ex-post. Our study builds on these ideas, extending them by designing an environment where we have more control over the informativeness of the qualification signals. This allows us to cleanly distinguish between explicit and implicit discrimination, which is not present in the previous literature. Second, our study provides a form of stress test of these ideas from social psychology using different methods. Given the recent replication crisis (Open Science

[8]We certainly do not do full justice to this extensive literature in our short summary in Appendix B. Therefore, we direct interested readers to the more comprehensive existing reviews such as Dovidio, Gaertner, and Pearson (2017).

Collaboration, 2015), it is reassuring to see that we find evidence that is consistent with the seminal work on aversive racism in the different context we consider, using different methods. Third, our experiment goes beyond the research in social psychology by giving incentives for accurate choices. Doing so allows us to demonstrate that social image concerns trump accuracy incentives. When thinking about the research in social psychology that is unincentivized, the omission of accuracy incentives may have implications for generalizing from the laboratory to field settings. Our findings help to alleviate such concerns. While the results from our study should also be interpreted with the relevant caveats associated with any laboratory study, our findings offer valuable insight by demonstrating the persistent nature of implicit bias even when incentives for accuracy are introduced.

In general, we view our study as an extension of the intellectual lineage emanating from the social psychology work on "aversive racism" by Joel Kovel, John Dovidio, and Samuel Gartner in the 1970s and 1980s. We contribute to the fledgling offshoot of this literature developing in economics. Given the recent appreciation within the economics literature of the critical role played by more complex and subtle forms of discrimination in generating disparities between groups, it seems important to include these ideas when trying to reach a more complete understanding of the various different forms that discrimination may take.

3 The Experiment

We administered an experiment consisting of two parts, each conducted with a separate group of participants. In the first part, the Job Candidate Assessment, we collected information from 80 participants. This included assessing their performance in several tasks – a general knowledge quiz (which we refer to as the *knowledge task*), a word search puzzle (which we refer to as the *word task*), and a matrix logic exercise (the job task, which we refer to as the *logic task*). Each of the participants also self-reported the gender that they identify with. These individuals were evaluated as *job candidates* by 240 participants serving as *employers* during the main part of the experiment, the Hiring Experiment, which we describe first.

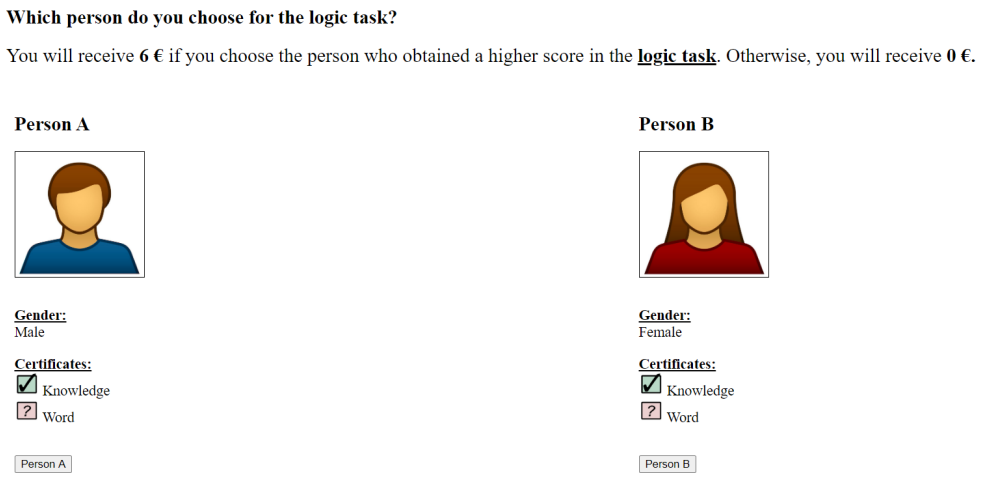
3.1 The Hiring Experiment

In the Hiring Experiment, employers went through a sequence of nine binary hiring decisions in which they decided which of two job candidates to hire (see Figure 1 for an example of the screen participants saw, translated to English). In each decision, employers were rewarded when they hired the candidate who performed better in the job task (with ties broken randomly). Across decisions, we systematically varied the CVs of the two candidates.

The CV of a candidate included three pieces of information: the gender of the candidate, and information about two possible qualifications that the candidate may or may not have. This qualification information was provided in the form of a *knowledge certificate* and a *word certificate*, which would certify that a candidate scored in the top 30% (i.e., top 24 out of 80)

in the word or knowledge task, respectively. For each of these qualifications, the CV either indicated that the candidate possessed the respective qualification with certainty (the green tick in Figure 1) or that it was unknown whether the candidate possessed that qualification (the question mark in Figure 1).[9] For instance, selecting a female candidate with a knowledge certificate therefore corresponded to selecting a random draw from all female candidates that scored in the top 30% in the knowledge task, irrespective of whether they also scored in the top 30% in the word task or not.

Figure 1: Screen in the HIRING EXPERIMENT.



Notes: (i) The figure displays a replication of an example screen that an employer would face in the experiment. (ii) The placement (left or right) on the screen of candidates was randomized. (iii) This example corresponds to "gender decision" Gen_K faced by employers. (iv) The gendered icons shown in the figure were identical for all CVs of the respective gender.

The nine hiring decisions faced by employers consisted of: one *complex* decision, where the choice was between a male and a female candidate who were *differently qualified*, two *gender* decisions between *equally qualified* male and female candidates, two *certificate* decisions between *differently qualified* candidates of the same gender, and four *simple* decisions between a male and a female candidate, where one candidate was *more qualified*.[10] Table 1 summarizes the decisions.

The experiment consisted of two between-subject treatment conditions, and a central part of the analysis exploits the within-subject dimension generated by the different hiring decisions that employers are presented with in sequence. The between-subject treatments only differed in the complex decision, which was the first decision to be taken. In treatment T_{fW}, the female candidate had a word certificate in the complex decision, and the male candidate had a knowl-

[9]This way of providing information leaves employers with maximal wiggle room to imagine how well candidates performed on the qualification task where it is uncertain whether they have the certificate or not. In forming this belief, any gender stereotypes they hold may influence their evaluation. In addition, it captures the idea that individuals tend to put positive information on their CV and that in the real world, evaluators may interpret the absence of a signal as a neutral, rather than negative, signal.

[10]Our terminology of differently/equally/more qualified refers to a gender-blind benchmark, i.e., comparisons where the gender of both candidates is ignored.

edge certificate. In treatment T_{fK}, the certificates were reversed in the complex decision, with the female candidate possessing a knowledge certificate, while the male candidate had a word certificate (see Table 1). The rationale for introducing this between-treatment variation was to allow us to examine whether participants shift their perception of the value of different certificates as a function of whether the male or female candidate held a given certificate (without inducing a demand effect for consistency by including these two decisions in sequence).[11]

Table 1: CVs of candidates in all nine hiring decisions in the experiment.

	Candidate A		Candidate B		
Label	**Gender**	**Certificate**	**Gender**	**Certificate**	
Com_{fW} (T_{fW})	f	W	m	K	Complex decisions
Com_{fK} (T_{fK})	f	K	m	W	
Gen_K	f	K	m	K	Gender decisions
Gen_W	f	W	m	W	
Cer_f	f	W	f	K	Certificate decisions
Cer_m	m	W	m	K	
Sim_{fK}	f	K	m	–	Simple decisions
Sim_{fW}	f	W	m	–	
Sim_{mK}	f	–	m	K	
Sim_{mW}	f	–	m	W	

Notes: Letter f refers to female candidate, m refers to male candidate, W refers to the word task, while K refers to the general knowledge task. Labels T_{fW} and T_{fK} distinguish the two treatments by referring to the respective CV of the female candidate in the complex decision. The labels A and B for the candidates are arbitrary, since they only represent the order in which they were presented on the screen, which was randomized.

In each hiring decision, the employer chose between two of the candidates who participated in the Job Candidate Assessment.[12] Importantly, the Job Candidate Assessment was completed earlier, and the employers' hiring decisions had no influence on the candidates' payoffs, nor did the candidates ever learn the employers' decisions. This was known by the employers. This feature of the design serves two purposes. First, it prevents the hiring decision from influencing the performance in the job task (e.g., in the spirit of a gift exchange). Second, it means we can rule out classical taste-based discrimination when interpreting our results.[13]

[11]An alternative design choice would have been to include both of these complex decisions in a within-subject design. However, this would have risked raising the salience of the experiment being about gender discrimination in the second complex decision, which would be identical to the first, with only the gender of the candidates reversed. Therefore, the salience of gender would be raised in a decision where we want the notion of gender discrimination to be obscured from the participant. While it is true that the later decisions, such as the gender decisions, may also raise the salience of gender, this is less problematic, since these decisions are intended to measure explicit discrimination.

[12]More specifically, the employer made a choice between candidates with the two CVs that they saw. Hence, each candidate corresponded to a random draw from all candidates with that CV; e.g., candidate "A" in figure 1 is a random draw from all male job candidates that scored in the top 30% on the knowledge task. The labels "A" and "B" were arbitrary and were randomized, along with the placement on the screen (left or right) of the two candidates.

[13]Aside from the fact that employers' hiring decisions do not affect the outcomes of the job candidates in our experiment, we also rule out the possibility that employers will need to interact with the candidate that they hire. We therefore rule out, by design, these two classical sources of taste-based discrimination.

To ensure that the employers had a good understanding of exactly what all the tasks completed by the job candidates involved, they were provided with printed sheets which showed the full tasks (i.e., the knowledge task, word task and logic task) that job candidates worked on in the Job Candidate Assessment.

Employers earned 6€ if they hired the better candidate in the complex decision and an additional 6€ if their candidate choice in a randomly drawn decision from all others was correct. At no point in the sequence of their nine hiring decisions did employers receive any feedback on their decision, and all payoff information came at the very end of the experiment. After each decision, employers had the option to sell their choice for 0.10€. Doing so meant that their initial hiring choice was replaced by a random draw from the two candidates.[14] We implemented this two-step procedure to gain greater insight into employers' motives. The initial step forces employers to rank one candidate above the other and thereby reflects hiring decisions in the real world, where one needs to make a concrete choice between distinct options. The second step measures employers' desire to actually implement the initial choice that they made between the candidates. An employer may choose not to do this if they believe the two candidates are equally good and therefore are indifferent about the choice of candidate, or when they wish to act "as if" they believe the two candidates are equally good.

To summarize, our incentive structure achieves multiple objectives. First, the fixed payment (as opposed to paying the employer in proportion to the candidate's output) ensures that the employer is incentivized to simply choose the candidate they believe is most likely to have performed better in the job task; thus, we remove the role of risk preferences as well as avoiding an excessive influence of high- (or low-) performing outliers. Second, the selling option allows employers to indicate that they believe the two candidates are equally likely to be the better candidate. Third, the hiring decision is inconsequential for the candidates, who never learn about the decision nor have their pay affected by it, which allows us to focus on belief-based forms of discrimination. By design, this rules out any gender discrimination based on concern for others' payoffs (material or psychological) or based on transactions with certain groups, and it differentiates our work from other experimental work on gender discrimination such as that of Bohren et al. (2022) or Reuben et al. (2014), where the evaluator's/employer's choices directly affect the candidate's fate.[15]

The order of decisions was partially randomized at the individual level. The treatment-specific complex decision was always taken first (either Com_{fW} or Com_{fK}). It was followed by a block of four decisions, randomly ordered, which are the gender and certificate decisions, and then another block of four decisions, randomly ordered, which are the simple decisions. This

[14] Expected-payoff maximization implies one would sell only if the subjective probability that one's initial choice is better lies below $0.5 + \frac{1}{60} = 0.5167$.

[15] While these features of the hiring choice may, at face value, seem somewhat artificial with respect to actual labor market discrimination, it is quite plausible that many hiring committees include members that (i) have an incentive (potentially intrinsic) to hire the best candidate for the organization, (ii) won't ever interact with whoever gets hired (or not hired) after the completion of the hiring process, and (iii) do not view it as their role to care directly about the candidates' outcomes. If gender enters such a committee member's evaluation, we would still call this gender discrimination.

partial randomization was implemented to limit the potential influence of order effects while ensuring that relatively more important decisions for our analysis appeared earlier. After the (initial) complex decision, we also measured the employers' beliefs about the informativeness of each of the two certificates about performance in the job task.[16] The reason for doing this was to assess whether participants shifted their beliefs in order to justify their complex decision's hiring choices as being qualification-based. For example, an employer holding a gender bias in favor of the male candidate, who happened to have a word certificate in the complex decision (i.e., who was presented in randomized treatment T_{fK}), might adjust their belief about the informativeness of the word certificate for job performance upwards to justify choosing the male candidate.

3.2 The Job Candidate Assessment

This Job Candidate Assessment was carried out prior to the main Hiring Experiment and served three purposes. First, it allowed us to run the Hiring Experiment with candidates drawn from the same subject pool as the employers, and to populate the candidates' CVs with real qualification data (as opposed to constructing fictitious candidates). This added to the realism of the task and allowed us to incentivize choices, including belief reports. Second, we were thus able to evaluate the decisions of employers against the true distribution of performance in the job task, i.e., to test the accuracy of the employers' revealed beliefs. Third, it provided us with an additional sample of participants, separate from the employers, from whom we could elicit beliefs about the association of gender with performance in the different tasks, to measure potential stereotypes present in the study population.

These job candidates completed multiple tasks and were scored on their performance in each. After the completion of all tasks, one of the tasks was randomly drawn to be paid out, with the participant's payoff linearly increasing in their performance. After they completed the tasks, we also elicited job candidates' beliefs about the performance of male and female candidates in the job task, as well as second-order beliefs. A more detailed description of the tasks and procedures in the Job Candidate Assessment experiment is provided in Appendix A.

[16] This was done in the following way. Employers were told that a candidate had been randomly chosen from the pool of candidates and they would earn 3€ if the candidate was in the top 50% in terms of performance on the job task. The employer was then given the option to replace this candidate with one who had a word or knowledge certificate, respectively, but would have to make a payment to do so. We elicited participants' willingness to pay to replace the randomly drawn candidate with a candidate holding a certificate for prices between 0.10€ and 1€ (in steps of 0.10€) in multiple price lists. 11% of participants made inconsistent choices in at least one of the lists (i.e., they switched multiple times). After the price list task, participants were also asked to indicate how informative they thought each of the two certificates was about performance in the job task on two non-incentivized 5-point Likert scales, with options ranging from "not informative" (1) to "very informative" (5). This provided a simpler, secondary instrument to measure essentially the same beliefs. We included the second measure due to the frequent miscomprehension and hence loss of observations in multiple price list tasks (Yu, Zhang, and Zuo, 2021). Our secondary belief measure resembles the common rating scales used in closely related psychology research (e.g., Uhlmann and Cohen, 2005).

3.3 Implementation details

We conducted 10 sessions of the HIRING EXPERIMENT with 24 participants each. Therefore, 240 participants took part in the experiment as employers, of which 119 (49.6%) were female. The average age was 24.5 years (SD: 4.9 years) and the majority were students in a STEM (52%) or Economics/Business program (33%). An additional and separate group of 80 participants participated in the JOB CANDIDATE ASSESSMENT, which comprised 4 sessions of 20 participants each, of whom 44 (55%) were female. The sessions were conducted between November 2017 and January 2018 at the WZB-Technical University laboratory in Berlin. Participants were invited to participate in the experiment using ORSEE (Greiner, 2015). The experiments were implemented in oTree (Chen, Schonger, and Wickens, 2016), and randomization into treatment in the HIRING EXPERIMENT was unconstrained on the individual level, resulting in 118 employers in Treatment T_{fW} and 122 employers in Treatment T_{fK}. Demographics by treatment assignment are reported in Table 8 of Appendix G.

4 Hypotheses

In this section, we explain how our experimental design allows us to identify both implicit and explicit discrimination from the choices made during the experiment. This allows us to test a series of hypotheses regarding the presence of different forms of discrimination in our data.

Since our study extends beyond the "rational" economic theories of taste-based or statistical discrimination, it is worthwhile clarifying what we mean by discrimination: An individual employer discriminates if their choice is causally affected by the candidates' gender. One way that we can see this in our experiment is by comparing pairs of choices, where only the gender of candidates is varied. For instance, suppose that an employer hires a male candidate with a knowledge certificate K over a female candidate with a word certificate W in the complex decision, Com_{fW}. This constitutes gender discrimination (against women) if this same employer would have chosen differently—instead favoring the candidate with the word certificate W—if the gender of the two candidates were reversed. This is precisely the decision employers face in the companion complex decision in our experiment, Com_{fK}. Since our experimental design rules out taste-based discrimination, an employer that would choose the male candidate in both these scenarios reveals that they believe that women are less competent than men in the job task, irrespective of which qualification they hold. Therefore, this behavior constitutes *belief-based discrimination*.

Our main interest in this project lies in better understanding discrimination in complex decisions, where the candidates under consideration are both "well-qualified" but their qualifications cannot easily be ranked. This scenario captures the essence of the final stage of many real hiring scenarios, where neither candidate is objectively superior to the other across every area of their professional profile. In such decisions between candidates with different strengths and weaknesses, either choice could be deemed non-discriminatory and be justified based solely on

qualifications. When this is the case, there is scope for a (conscious or unconscious) discriminatory bias to shift the weight placed on different qualifications when evaluating candidates, thereby favoring the candidate of a particular gender. In the context of our experiment, candidates are being hired to perform a logic task where there is a prevailing stereotype that men tend to perform better on average.[17] We hypothesize that this gender stereotype may lead to discrimination against women. Importantly, this discrimination may manifest in different ways—employers may consciously and intentionally favor male candidates, engaging in *explicit* discrimination, or they may only discriminate when they can hide it from themselves and/or others, thereby engaging in *implicit* discrimination.

4.1 Aggregate gender bias in the complex decisions

While a core objective of our study is to detect the presence of these different forms of discrimination, as a first step, we examine the aggregate level of discrimination in the complex decisions. We do so using a between-subject approach that is analogous to the method used to identify framing effects (see Tversky and Kahneman, 1981, for a classic example). This involves comparing the hiring rates of men and women in the pair of complex decisions that are identical other than the gender of the two candidates being reversed.[18] To illustrate how this allows us to measure aggregate discrimination, consider the following simple example. Suppose that in the first complex decision, Com_{fW}, the male candidate is hired by 45% of employers, while 75% hire the male candidate in Com_{fK}. This implies that men are hired at an average rate of 60% across the two decisions. This implies that gender information is causally influencing decisions and reveals an aggregate gender bias in favor of men of $(60-40)\% = 20\%$. Importantly, this measurement of bias provides a *lower* bound on the fraction of employers that would choose the male candidate in *both* decisions, and hence on discrimination against women.[19] The lower of the two hiring rates of 45% provides an upper bound. While this is a fairly standard application of a between-subject analysis, in Appendix C we provide a formal theoretical discussion of the relationship between these measurement concepts, namely the observed aggregate (male) gender bias in a population and the prevalence of individual

[17]Prior to running the HIRING EXPERIMENT, we elicited first- and second-order beliefs of the job candidates about the performance of men and women in this task. While the actual performance of men and women was approximately equal on average, job candidates believed that men tend to perform better in this task. In Appendix F, we provide a more detailed discussion of this statistically inaccurate belief. It is in line with the classical stereotype according to which analytical and logical reasoning are rather "male" abilities (e.g., Heilman, 2012).

[18]If we adopt a normative benchmark of non-discrimination (in the sense that gender information should not be used at all to guide hiring decisions), then information about gender can be thought of as a frame of the qualifications. Employers may value a qualification differently depending on whether it is held by the male or female candidate. It is for this reason that we opted to use a between-subject design for our complex decisions. Employers may be able to convince themselves that the qualification held by the male candidate is more valuable if they face a single complex decision, but not if they face two near-identical decisions with only the gender of the candidates reversed.

[19]The male bias yields a *lower* bound, because there may be some employers that discriminate in favor of women. The male bias is the *net* effect. This reflects many measurements of aggregate bias in everyday life, which capture the bias against women net of any positive discrimination (affirmative action) in favor of women.

discrimination (against women). Using this approach, we will evaluate the following testable hypothesis.

Hypothesis 1 (Aggregate Gender Bias)**.** *In the complex decisions, male candidates are hired more often. This reflects a bias towards hiring men and reveals a lower bound on discrimination against women.*

4.2 Comparing gender bias in the complex and gender decisions

Having tested for the presence of an aggregate gender bias in Hypothesis 1, we wish to establish whether part of this discrimination in the complex decisions is due to *implicit discrimination*. We define an implicit discriminator as an individual who only engages in discriminatory behavior in situations where their actions do not explicitly reveal their bias to themselves or others. A decision-maker who engages in implicit discrimination does not consciously exhibit explicit discriminatory behavior. However, they do discriminate when they can plausibly remain unaware of their own biased actions. A decision-maker who discriminates *explicitly* will be willing to discriminate in both types of scenarios—those where the discrimination is opaque and those where it is clearly revealed by choices.

We will examine the influence of implicit discrimination in our data in two ways. First, we ask whether it leads to a shift in the net discrimination against women when comparing the complex decisions with the gender decisions. Second, we exploit the full set of hiring decisions in our experiment to detect implicit discrimination by individuals.

In our experiment, we view the complex decisions as capturing scenarios where discrimination remains opaque. In such situations, an employer's choices may be influenced by their (unconscious) stereotypes or biases. In contrast, we capture scenarios where discrimination is obvious or explicit by means of the gender decisions, Gen_W and Gen_K, in our experiment. These decisions involve hiring choices where both candidates are identically qualified and differ only in their gender. From the "gender-blind" perspective of non-discrimination, any strict non-discriminating employer should be indifferent. We allow them the opportunity to post hoc express indifference by offering them the option to "sell" their choice. Therefore, individuals that are willing to discriminate explicitly will discriminate in the gender decisions and the complex decisions; individuals that are not willing to discriminate explicitly, but hold an unconscious gender bias will only discriminate in the complex decisions.

Implicit biases tend to be those that are suppressed due to being in conflict with a widespread normative prescription in society. One such norm in modern Western society is to not discriminate against women.[20] Therefore, one might expect employers who hold a bias against women to be less willing to act on it in the clear-cut gender decisions than in the more opaque complex decisions. While this mechanism operates at the level of the individual, if it

[20]Another prevalent norm is the norm of not using gender information at all. An individual that views this as the more important norm would display indifference in the gender decisions, but then reveal their bias in the complex decisions.

exerts a dominant effect, it could affect the aggregate level of net bias observed when comparing the gender and complex decisions.[21] We test this in Hypothesis 2. The male bias for the gender decisions is calculated in the same way as in complex decisions.[22]

Hypothesis 2 (Gender Decisions vs. Complex Decisions)**.** *The male bias in hiring is lower in the gender decisions than in the complex decisions.*

4.3 Detecting implicit discrimination by individuals

Hypothesis 2 provides valuable insight into how the net gender bias changes with the complexity of hiring choices. However, this comparison does not fully utilize the rich within-subject information collected in our experiment, nor does it allow us to evaluate what fraction of individuals discriminate implicitly.[23] To detect implicit discrimination, we wish to identify employers who favor one gender in the explicit gender choices and the other in the more opaque complex choice. We, therefore, first use the two choices in the gender decisions of every employer to classify them into one of several *explicit discrimination types*. Given our primary focus on implicit gender discrimination against women, our main interest lies in examining employers who do not engage in explicit discrimination against women. It is important to note that if an employer is willing to discriminate explicitly against women, they cannot, by definition, engage in implicit discrimination against women.[24] We detect implicit discrimination against women in our experiment when an employer is willing to discriminate against women in the complex decisions, but does not discriminate against women in the two gender decisions where discrimination is explicit.

Hypothesis 3 (Explicit and Implicit Discrimination)**.** *To detect implicit discrimination, we consider the set of employers that do not discriminate explicitly against women. This includes (a) those who discriminate explicitly against men, and (b) those who engage in explicit non-discrimination. Among each of these two groups of employers, a majority choose to hire the male candidate in the complex decision, thereby displaying implicit discrimination against women.*

[21] To do so, it would need to dominate other mechanisms that may operate when moving from the the clear-cut gender decisions to the more opaque complex decisions. For example, it could be offset by individuals who comply with a strict non-discrimination norm in the gender decisions, but then display affirmative action (positive discrimination) in the complex decisions.

[22] In addition to calculating the average male bias across the two gender decisions, we can calculate the bias within each decision separately. Here, there is a clear non-discrimination benchmark of hiring male and female candidates at equal rates. So, for example, we can calculate a male bias for the decision Gen_K by comparing the hiring rate of men and women in that hiring decision. If in Gen_K men (vs. women) are hired at a rate of 55% (vs. 45%), then it has a male bias of $(55-45)\% = 10\%$.

[23] As noted in footnote 21, there are different mechanisms that may be triggered when moving between the complex and gender decisions. Hypothesis 2 evaluates the net effect of all of these mechanisms.

[24] Our main interest is in implicit discrimination against women because implicit discrimination is likely to manifest when an individual's implicit preferences are in violation of a social norm. This is the case for discrimination against women, where it violates a strong social norm. Since the social norm against positive discrimination (affirmative action) in favor of women is arguably weaker, we would expect less implicit discrimination against men. Nevertheless, the methodology used here is portable and can be used to study different forms of implicit discrimination.

It is important to note that such implicit discrimination is distinct from (accurate and inaccurate) statistical discrimination. Under statistical discrimination, employers hold beliefs about the informativeness of gender for the outcome of interest (here, performance in the job task) and, therefore, use gender as a signal. When employers display implicit discrimination, the choices they make in different decisions cannot be rationalized by holding stable beliefs about the informativeness of gender as a signal about performance. The choice of the female candidate in the gender decisions indicates that the employer views the signal "female" as a positive signal about performance in the job task, while they express the opposite belief by choosing the male candidate in the complex decisions. Indeed, this hypothesis operationalizes the signature prediction of the general framework of implicit preferences proposed by Cunningham and de Quidt (2023), using a between-subjects design for the main decisions of interest in which implicit preferences are suspected to matter and extending this by using the within-subjects dimension to elicit explicit preferences.[25]

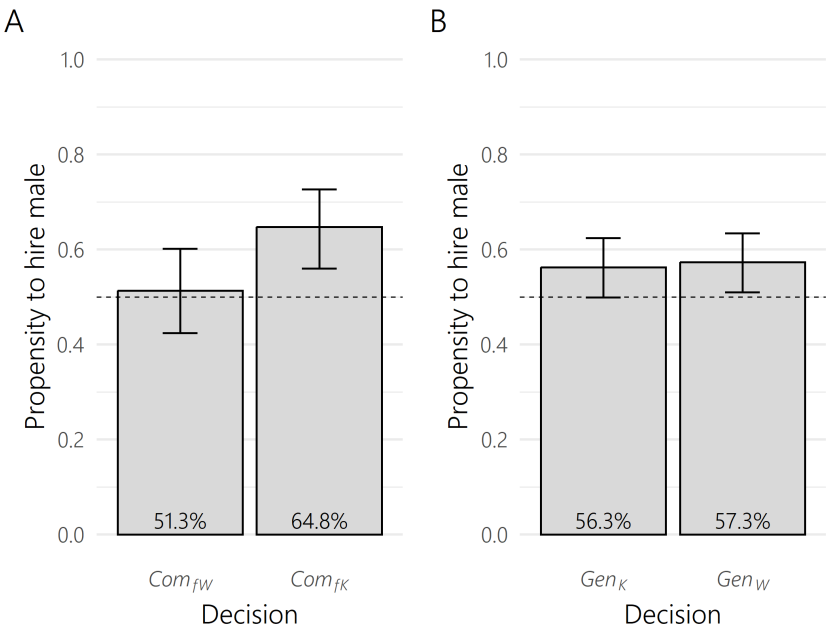
In our analysis, in addition to studying the implicit discrimination of those who discriminate explicitly against men, we can also analyze the behavior of those who engage in explicit non-discrimination to reveal implicit preferences. This allows us to evaluate whether this group who consciously comply with the norm of non-discrimination displays the hypothesized implicit bias against women.

4.4 A comment on statistical tests

In the analysis below, we will often be comparing the hiring rates of men and women. In doing this, we evaluate the null hypothesis that there is no gender bias (i.e., that male candidates are hired with a probability of 0.5) against the two-sided alternative using Pearson's χ^2 test. For cases when multiple decisions of the same employer are included in one comparison, dependence of observations is taken into account via using the Cochran-Mantel-Haenszel test statistic instead of the standard Pearson χ^2 statistic (Agresti, 2013, p. 227). For inferring a non-zero, positive lower bound of implicit discrimination, a one-sided test is used with the null hypothesis that the lower bound is zero or less (distinguishing between a zero and a negative lower bound is not meaningful for identifying implicit discrimination in a given group of explicit discriminators). The reported 95% confidence intervals for proportions are Wilson score intervals.

[25]The authors also show how this behavioral prediction—which they call the "figure-8" pattern—unifies various theories of implicit preferences, which in our setting corresponds to an implicit belief bias favoring men that unconsciously causes discrimination against women.

Figure 2: Ultimate hiring choices in complex (A) and gender (B) decisions.



Notes: (i) Panel A reports the propensity to hire the male candidate in the two complex decisions. (ii) Panel B reports the propensity to hire the male candidate in the two gender decisions. (iii) Error bars show 95% confidence intervals. (iv) The dashed horizontal line indicates the non-discrimination benchmark, namely hiring men and women at the same rate.

5 Main Results: Explicit and Implicit Discrimination

5.1 Gender bias in complex and gender decisions

Figure 2 reports the average propensity of employers to hire the male candidate in each of the complex and gender decisions. The left panel considers the complex decisions. Here, the two candidates are *differently qualified*. The bars denote the employers' ultimate decisions after the option to sell has been exercised. When calculating these ultimate propensities, each choice that is sold contributes equally to both candidates. This means that a probability of 0.5 is assigned to each candidate for a choice that is sold.[26] We see that when the two candidates are differently qualified, employers display a significant bias in hiring in favor of men: When averaging across the two decisions, men are hired around 58.0% of the time in the ultimate decisions $[0.5 \times (51.3 + 64.8)\%]$, which is significantly larger than the gender-neutral 50% ($\chi^2_{\text{df}=1} = 6.2$, $p = 0.013$). This means that there is a significant fraction of employers that would hire the male candidate irrespective of which of the two qualifications he holds in the

[26]Therefore, if all employers opted to sell their initial choice, the male hiring propensity would be equal to 0.5. In Figure 11 in the Appendix, we reproduce this figure with both the initial and ultimate decisions. In Figure 11, the grey bars denote the employers' initial decisions, while the blue bars reflect their ultimate decisions. We see that the observed pattern of behavior is very similar irrespective of whether we use the initial decisions or the ultimate decisions.

complex decisions (i.e., in both Com_{fW} and Com_{fK}). In these two complex decisions, employers appear to view the word certificate as being a stronger signal of success in the job task, since we observe that women are hired at roughly the same rate as men when the woman holds the word certificate in Com_{fW}. In contrast, when the man holds the word certificate in Com_{fK} then men are hired at a much higher rate (close to two-thirds). These results provide evidence in support of Hypothesis 1.

Result 1 (Aggregate Gender Bias)**.** *In the complex decisions, male candidates are hired 58.0% of the time, which is significantly higher than the rate of hiring women of 42.0%. This indicates a clear bias in favor of hiring men, revealing a lower bound on discrimination against women of 16.0%.*

Turning attention to the right-hand panel, we see that in the gender decisions, we also observe a bias in hiring in favor of men. Here, the two candidates are *equally qualified*. Men are hired ultimately at a rate of 56.3% ($\chi^2_{\text{df}=1} = 3.75$, $p = 0.053$) and 57.3% ($\chi^2_{\text{df}=1} = 5.10$, $p = 0.024$) in Gen_K and Gen_W, respectively, yielding an average of 56.8%. This means that a significant fraction of employers are willing to hire a male candidate over an equally qualified female candidate even when this explicitly reveals a gender bias. One potential explanation is that employers hold a stereotype that men are better at the job task and are willing to express this stereotype overtly. In terms of evaluating Hypothesis 2, the overall male bias in the gender decisions (13.5%) is only slightly smaller than in the complex decisions (16.0%), so we do not find evidence in support of Hypothesis 2 ($\chi^2_{\text{df}=1} = 0.09$, $p = 0.77$).

Result 2 (Gender Decisions vs. Complex Decisions)**.** *The male bias is not significantly different between the gender decisions and the complex decisions. In the gender decisions, male candidates are hired 56.8% of the time, which is comparable to the hiring rate of men of 58.0% in the complex decisions.*

Our analysis of the aggregate-level results reveals a gender bias that favors men in both types of hiring decisions: those between equally qualified candidates and those between differently qualified candidates. However, these aggregate-level results are consistent with different explanations in terms of the type of discrimination that could be generating them.[27] This illustrates the importance of the within-subjects dimension of our design, which we now exploit

[27]For instance, one interpretation is that implicit discrimination is absent, with approximately 15% of employers explicitly discriminating by consistently selecting the male candidate, while the remaining employers consistently make gender-blind decisions based solely on certificate information. Another interpretation suggests high implicit discrimination, with 42% of employers consistently choosing the male candidate in complex decisions and the female candidate in gender decisions, while another 35% do the opposite. Among the remaining 23%, 9% consistently choose the male candidate, while 14% choose the male candidate in gender decisions and the candidate with the word "certificate" in complex decisions. Such a scenario involves 77% of employers discriminating implicitly. These examples illustrate how the aggregate-level results align with a wide range of interpretations regarding the underlying source of discrimination, highlighting the importance of the within-subjects dimension in our study design.

to shed light on the existence of implicit discrimination and its relation to explicit discrimination.[28]

5.2 Types of explicit discrimination, and implicit discrimination

To test for the presence of implicit discrimination, it is necessary to first assign individuals to explicit discrimination types. The reason for this is that, based on our definition, an implicit preference is only revealed when it is inconsistent with a preference that an individual is willing to display explicitly. To achieve this classification into explicit discrimination types, we use every individual employer's two choices in the gender decisions. Since both candidates are equally qualified in the gender decisions, when an employer chooses *not* to sell their choice, this reveals both that they: (i) believe gender matters here, and (ii) are willing to let this information affect their hiring decision. Accordingly, we define an employer as engaging in *explicit discrimination* if she opts not to sell her gender choice in at least one of the two gender decisions. If she sells both of her gender choices, we classify her as engaging in *explicit non-discrimination*. Moreover, we distinguish different types of explicit discrimination according to the decisions employers make: We say that an employer engages in *explicit discrimination against women [men]* if this employer chooses the male [female] candidate in both of their initial gender decisions. Explicit discriminators that cannot be categorized in this way because they choose the male candidate in one decision and the female candidate in the other will be said to engage in *mixed explicit discrimination*.[29]

To move from explicit to implicit discrimination, it is important to consider social norms. Implicit discrimination is likely to manifest when there is a tension between what is viewed as socially appropriate behavior and what an individual would do in the absence of the social environment.[30] Therefore, to relate our classification of explicit discrimination to the prevailing social norms around gender, those employers that do not discriminate explicitly (and only those) are complying with a strong gender-symmetric norm of non-discrimination that prescribes completely gender-blind decision-making. By contrast, the norm of not discriminating *against women* is also satisfied by those employers that discriminate explicitly against men. This norm is gender-asymmetric and slightly weaker in the sense that it allows for some discrimination in the form of "affirmative action favoring women." In particular, it allows for a hiring rule that endorses the preferential hiring of women over equally qualified men. This asymmetric norm is important to consider as it reflects a current feature of hiring practices

[28]For an overview of how all decisions are related and all our employers' aggregate choices in each treatment, see Figure 9 and Table 9 in Appendix G. They summarize aggregate behavior in terms of both initial and ultimate hiring, the latter being obtained from assigning equal probability to both candidates for every sold choice.

[29]There are several reasons why an employer may exhibit this behavior. For example, an employer who seeks to maximize their expected payoff may fall into this category if they believe that men outperform women on the job task when holding a word certificate, but women outperform men when holding a knowledge certificate.

[30]For clarity, when we refer to the "absence of the social environment" here, we do not simply mean that an individual's actions are not observed. Social norms may well be internalized and exert a strong influence on an individual's behavior even when they know that the only person who will observe their action is themselves.

in many societies—e.g., in Germany, South Africa, and other countries, where it is common for organizations to have an explicit policy to hire a female candidate over an equally qualified male candidate. Our classification into types considers each of these two types of explicit norm compliance separately. Each can generate implicit discrimination.

Table 2: Hiring in complex decisions by explicit discrimination type

Discrimination type	**Frequency of disc. type (%)**		**Propensity to hire male candidate (%)**	
	Treatment T_{fW}	Treatment T_{fK}	Decision Com_{fW}	Decision Com_{fK}
	(1a)	(1b)	(2a)	(2b)
Explicit discrimination	67.8	73.0	57.5	64.0
– against men	13.6	12.3	62.5	66.7
– against women	25.4	26.2	60.0	68.8
– mixed	28.8	34.4	52.9	59.5
Explicit non-discrimination	32.2	27.0	28.9	63.6

Notes: (i) Columns (1a) and (1b) report the frequency distribution of different discrimination types, by treatment (in %), based on the classification using the gender decisions. (ii) Columns (2a) and (2b) report the propensity to initially hire the male candidate in the complex decision for each discrimination type, by decision version/treatment (in %).

Table 2 reports the distribution of explicit discrimination types for each treatment, as well as the propensity of each type to initially hire the male candidate in each complex decision.[31] The (*a) columns report the distributions and hiring propensities for the treatment in which the female candidate holds the word certificate in the complex decision, while the (*b) columns correspond to the treatment in which the male candidate holds the word certificate. Columns (1a) and (1b) show that the type distributions are similar between treatments (there is also no evidence for a treatment effect on the type distribution, $\chi^2_{\text{df}=3} = 1.2$, $p = 0.75$). This is reassuring, since our type classification is only based on the (treatment-invariant) gender decisions. Approximately 70% of employers discriminate explicitly. Within this group, about twice as many discriminate against women (26%) as against men (13%), with the remainder classified as "mixed". Unsurprisingly, employers who are willing to discriminate explicitly against women in the gender decisions display a strong bias in favor of male candidates in the complex decisions, hiring men at rates of 60.0% in Com_{fW} and 68.8% in Com_{fK}. However, perhaps more surprisingly, we find that those employers that discriminate explicitly *against men* in the gender decisions exhibit a substantial bias *in favor of men* in the complex decisions. Strikingly, when comparing the group of individuals who discriminate explicitly against women with those who discriminate explicitly against men, there is a remarkable similarity in their propensity to hire the male candidate in the complex decisions. We see this in columns (2a) and (2b), where the propensities to hire a man are 62.5% vs. 60.0% in Com_{fW} and 66.7% vs. 68.8% in Com_{fK}.

[31]The rationale for using the initial decisions as an indication of the employer's strict preference in the complex decisions (and only in the complex decisions) is the following. The complex decision was always the first decision that an employer made. At the time of making this first hiring decision, the employer was not yet aware that they would be offered the opportunity to sell their choice. The opportunity to sell was only described after making the complex decision. Therefore, it is reasonable to assume that employers revealed their true (strict) preference through their initial decision in the complex hiring choice. In later hiring choices, one cannot rule out that employers who were close to indifferent strategically anticipated the opportunity to sell in their initial choices (since they had the opportunity in previous hiring decisions), which is why we focus on ultimate choices in the subsequent hiring choices. While we believe that using the initial choices for the complex decision is the most valid approach, to alleviate any residual concerns, in Appendix D we show that our results are robust to this analytical choice by using the ultimate decisions.

This large gender bias in favor of male candidates in the complex decisions among employers that discriminate explicitly against men identifies a strict notion of implicit discrimination. These employers comply with a norm of not discriminating against women in the gender decisions, where discrimination is overt, yet they do discriminate against women in the complex decisions, where their bias is obscured. Specifically, among the 13% of all employers who express an explicit bias against men via their choices in the gender decisions, over 60% choose to hire the male candidate in the complex decisions, implying that roughly between 30 and 60% discriminate implicitly against women (corresponding to 4–8% of *all* employers).[32] Testing for a positive lower bound indeed suggests the presence of implicit discrimination in this group ($z = 1.62$, $p = 0.052$). Thus, we find support for Hypothesis 3.

Result 3a (Implicit and Explicit Discrimination)**.** *We find that 13% of our employers discriminate explicitly against men. Of these, over 60% choose to hire the male candidate in the complex decisions. This implies that approximately 30 to 60% of them display implicit discrimination, with the lower bound being significantly larger than zero.*

As noted above, in the discussion of Hypothesis 3, the pattern of behavior identified here (between-subjects) cannot be rationalized by any transitive preferences. In particular, it cannot be rationalized by statistical discrimination for any (accurate or inaccurate) subjective beliefs.[33] This identification of the presence of implicit discrimination is our main finding, rather than the quantification of its level, which may vary depending on the population and specific features of the experimental design. However, it is noteworthy that within each of the two complex decisions, the propensity to hire a man is remarkably similar between employers who discriminate explicitly against men and those who discriminate explicitly against women. This is consistent with an interpretation where the distinction between these two types lies solely in whether or not they are willing to discriminate explicitly against women and thus visibly violate an anti-discriminatory norm. Under this interpretation, both types share a belief or stereotype that men are superior at the job task.

When considering the group of explicit non-discriminators, we find that they hire the male candidate 46.3% of the time in the complex decisions, averaging over treatments. Therefore, we estimate a lower bound of implicit discrimination against women of −7.4% in this group. The negative estimate indicates that we do not find evidence of implicit discrimination against

[32]Specifically, the average rate at which this explicit discrimination type hires the male candidate in a complex decision equals 64.6%=(62.5+66.7)/2%. The lower bound is then given by 29.2% = 64.6−(100−64.6)%. The upper bound is 62.5%, which is obtained from the smaller of the two hiring rates between Com_{fW} and Com_{fK}. See Appendix C for a formal derivation of the underlying formulas.

[33]Given our incentives, which let employers compare distributions, rational choices are guaranteed to satisfy transitivity if the same candidate profile is perceived as the same candidate draw across decisions (e.g., "the" female candidate with a knowledge certificate in complex decision Com_{fK} and gender decision Gen_K). If an employer perceives these as independent draws in each binary decision, rational choice may be intransitive, as with "intransitive dice" (see, e.g., Savage, Jr., 1994). In this case, a sufficient condition for transitivity would be, for instance, that all distributions/beliefs are normal distributions with identical variances but different means, so there is first-order stochastic dominance. Reassuringly, given actual performance distributions and the perception of independent draws, our candidate profiles would not give rise to intransitive choices (see Table 7 in Appendix F).

women among individuals who display a commitment to complying with the principle of non-discrimination.[34]

Result 3b (Implicit Discrimination and Explicit Non-Discrimination)**.** *We find that approximately 30% of our employers engage in explicit non-discrimination. Amongst these, we do not observe a statistically significant gender bias against women in the complex decisions. We, therefore, do not detect implicit discrimination against women in this group.*

5.3 Robustness exercises

To provide support for the validity of our main findings, we conduct a series of robustness exercises. These exercises, along with the associated results, are described in detail in Appendix D. The following is a brief overview of the exercises we conduct.

First, we evaluate the robustness of our results to conducting the analysis using the initial versus the ultimate hiring propensities from the complex decisions. We find that our main result is largely unaffected by which specification is used because very few of the employers in question sell their initial choices. Second, we consider two alternative approaches to classifying employers into explicit discrimination types. Under both approaches, in line with our main result, we detect a substantial amount of implicit discrimination. This analysis also provides evidence that our results are not driven by a "balancing heuristic" whereby employers try to balance their hiring portfolio by equalizing the number of men and women they hire across the sequence of decisions they make. Finally, to examine whether our results may be generated by noise or some other characteristic of our design, we conduct a set of placebo checks. To do this, we replicate our analysis by checking for implicit discrimination (i) against men, (ii) against the word certificate, and (iii) against the knowledge certificate. We fail to detect implicit discrimination in any of these three placebo exercises, providing evidence that our main results are not driven by noise or some other design feature.

6 Further Results

6.1 An additional test for implicit discrimination

Our final robustness check indicates that any implicit bias in decision-making pertains solely to gender, not the two types of certificates (knowledge and word). Under this assumption, we can attribute any preference reversal over certificates when comparing a complex decision (different gender) and a certificate decision (same gender) to gender information.[35] This is

[34]Although Hypothesis 3b is about implicit discrimination against women, we can also use the data to test for implicit discrimination more generally among explicit non-discriminators with a two-sided test. There is no evidence for it ($z = -0.62$, $p = 0.53$). See also Appendix D for inferences about implicit discrimination against other attributes than female gender.

[35]When we refer to a preference reversal over certificates, we mean, for example, that an employer prefers the candidate with the word certificate in a complex decision, but prefers the candidate with the knowledge certificate in a certificate decision.

because, if we remove gender information, the complex decisions and certificate decisions are all identical to one another—they all involve choosing between one candidate with a word certificate and one with a knowledge certificate.

In particular, this assumption allows us to detect implicit discrimination by individuals in a different way. To see this, consider an employer who chooses a male candidate with a knowledge certificate over a female candidate with a word certificate in the complex decision—i.e., $(m,K) \succ (f,W)$. Suppose that this employer reverses their preference over certificates in one of the certificate decisions, say by exhibiting $(f,W) \succeq (f,K)$ in decision Cer_f. Finally, suppose that the employer also reveals a weak preference for the female candidate in the gender decision with knowledge certificates—i.e., $(f,K) \succeq (m,K)$. Such an employer reveals a (weak) explicit preference for women in the gender decision. However, in the other two decisions, they reveal that they adjust their valuation of the certificate across decisions, valuing it more when the male candidate holds it. This indicates implicit discrimination since the employer displays no bias towards men when gender discrimination is explicit, but a bias towards men when it is implicit. This pattern cannot be rationalized by any single transitive preference (or set of consistent beliefs) over candidates. Furthermore, there are three other similar non-transitive cycles possible within our experimental design—in each of the two treatments, there is one involving the male certificate decision, Cer_m, and one involving the female certificate decision, Cer_f.[36]

To pin down implicit discrimination using this approach, we consider the set of all employers who reveal at least a weak preference for women on both gender decisions—this includes all employers who explicitly discriminate against men as well as those who explicitly do not discriminate. These are employers who reveal that $(f,W) \succeq (m,W)$ and $(f,K) \succeq (m,K)$. Among these employers, we compute the frequency of strict certificate preference reversals between the complex and both certificate decisions, since this identifies implicit discrimination.[37] We find that 42.5% of our sample reveals at least a weak explicit preference in favor of women in the gender decisions (this can also be calculated by adding the type frequencies in Table 2). Averaging over treatments, 22.8% of these choose the man in the complex decision and exhibit a reversal of their certificate preference between the complex and certificate decisions, implying that they discriminate implicitly against women. This amounts to 9.6% of our entire sample. This estimate is larger than the upper bound of around 8% we obtained from our main analysis. The reason for this is that the classification used in our main analysis only looked for implicit discrimination amongst those who explicitly discriminate against men. Here, we also allow for the detection of implicit discrimination against women amongst those who don't discriminate explicitly.

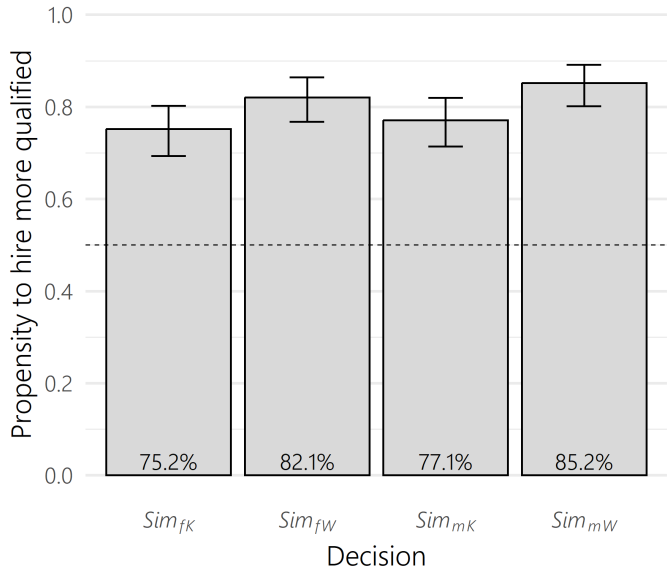
[36]These correspond to the "right triangles" in Cunningham and de Quidt (2023). Specifically, the cycle described in the main text corresponds to $(f,K) \succeq (m,K) \succ (f,W) \succeq (f,K)$. The other three are $(m,W) \succeq (m,K) \succ (f,W) \succeq (m,W)$, and $(f,W) \succeq (m,W) \succ (f,K) \succeq (f,W)$, and $(m,K) \succeq (m,W) \succ (f,K) \succeq (m,K)$.

[37]Recall that we always interpret the initial choices in the complex decision as revealing a strict preference. The rationale is further explained in the table notes below Table 2. In all other decisions, strict preference additionally requires to not sell the choice after it was made.

6.2 Simple decisions

In this section, we consider the simple hiring decisions, all of which are between one candidate with a certificate (knowledge or word) and another without a certificate. The two candidates being compared always differ in their gender. We refer to the candidate with the certificate as being *more qualified*. Studying these hiring decisions allows us to evaluate whether employers view the certificates as informative about performance in the job task. It also allows us to assess whether we observe a gender bias in a hiring context where one candidate is *more qualified*. We have already seen above that there is a gender bias in favor of men in scenarios where the two candidates are *differently qualified* (complex decisions) or *equally qualified* (gender decisions).

Figure 3: Ultimate hiring choices in simple decisions.



Notes: (i) The figure shows the propensity to hire the more qualified candidate in the four simple decisions, where the gender and certificate of the more qualified candidate varies. (ii) As described in Table 1, in decisions Sim_{fK} and Sim_{fW} the more qualified candidate is female, with a knowledge certificate in the former and a word certificate in the latter; analogously, in decisions Sim_{mK} and Sim_{mW} the more qualified candidate is male, with a knowledge certificate in the former and a word certificate in the latter. (iii) In all four decisions, the comparison is between a male and a female candidate. (iv) Error bars show 95% confidence intervals. (v) The dashed horizontal line indicates an equal aggregate propensity to hire both candidates in a particular choice.

Figure 3 reports the average propensity to hire the more qualified candidate in these four simple decisions. (Figure 12 in the Appendix reproduces this figure with both the initial and ultimate choices.) Moving from left to right, the figure first shows the hiring propensities in the two decisions where the female candidate is more qualified (holding a word or knowledge certificate). Thereafter, the four right-most bars show the hiring propensities where the male candidate is more qualified.

In all four decisions, the vast majority of ultimate hiring choices favor the more qualified candidate: Irrespective of whether that candidate is female or male, she/he is hired at a rate

of 75–85%. This shows that employers view the certificates as being informative predictors of performance in the job task, irrespective of the gender of the candidates. However, it is clear from the figure that the more qualified candidate is not always hired—in approximately 20% of these decisions, the employer opted to hire the candidate without the certificate. We, therefore, check whether there is a gender bias in hiring in these simple decisions. While the direction of the estimated gender bias suggests that more qualified women are chosen slightly less often than more qualified men, the bias is small and not significant ($\chi^2_{\text{df}=1} = 0.92$, $p = 0.34$). The magnitude of the bias is 1.9% in the "simple knowledge decisions" (comparing Sim_{fK} and Sim_{mK}) and 3.1% in the "simple word decisions" (comparing Sim_{fW} and Sim_{mW}).

6.3 Redefining merit

In this section, we explore the idea that hiring decisions between candidates that are *differently qualified* presents employers with an opportunity to hide or obscure their discrimination. An employer who holds gender-biased beliefs faces a decision problem that is psychologically very different in the complex decisions in comparison to the gender decisions. In the complex decisions, the different qualifications of the two candidates provide the employer with the opportunity to make a gender-biased decision, while convincing themselves that this decision is based solely on the candidates' qualifications, i.e., to justify discrimination by "redefining merit" (Uhlmann and Cohen, 2005). For example, an employer may (subconsciously) convince themselves that whichever certificate the male candidate has is the certificate that is most informative for predicting performance in the job task. In contrast, in the gender decisions, this is not possible—an employer that makes a gender-biased hiring decision here must confront their bias directly.

We are able to investigate this psychological channel here by comparing employers' beliefs about the predictive value of each of the two certificates (knowledge, word) for the job task between the two treatments T_{fW} and T_{fK}. These beliefs are elicited immediately after the complex decision, which is the first hiring decision that employers complete. In one of our two treatments, T_{fK}, the female candidate has the knowledge certificate and the male candidate has the word certificate in the complex decision. The certificates are reversed in the other treatment, T_{fW}. This implies that we can compare the beliefs about the predictive value of the certificates in a between-subject comparison to assess whether we find evidence of redefining merit. If some participants justify their discrimination against women in the complex decisions by redefining merit, then employers' judgment of the word certificate's value relative to that of the knowledge certificate should be more favorable in Treatment T_{fK}, where the male candidate has the word certificate than in Treatment T_{fW}, where the female candidate has the word certificate.

In Appendix E, we provide a detailed discussion of the analysis of certificate valuations across the two treatments. These results indicate that employers are redefining merit by adjusting their valuation of the certificates in a manner that is consistent with trying to justify

a male bias. The relative valuation of the word certificate against the knowledge certificate is increased by estimated 0.09 € in the treatment T_{fK}, where it is presented on the male candidate's CV. We provide additional support for this result by showing that the effect is particularly pronounced amongst the set of employers who discriminate explicitly by gender.

6.4 Statistical (non-) discrimination

A key distinguishing feature of our study is that our identification of discrimination does not rely on any assumptions regarding the employers' subjective beliefs. Since (accurate) statistical discrimination serves as an important benchmark in much of the literature, however—see Bohren et al. (2022)—we include an extensive discussion of our findings relative to this benchmark in Appendix F. The main insights from this discussion are the following. First, men and women were on average equally productive in the job task. Second, both gender information and the certificates are informative signals for hiring decisions. For example, an employer who engaged in accurate statistical discrimination based on the true statistical frequencies in our sample would be classified as displaying "mixed" explicit discrimination in the gender decisions, hiring the female candidate in Gen_K and the male candidate in Gen_W (without any selling). However, not a single employer's choice pattern is consistent with accurate statistical discrimination across all nine decisions, meaning that every employer who is discriminating by gender in our experiment is statistically inaccurate to some extent.

The main lesson we draw from these observations is that while assuming statistically accurate beliefs may be an appropriate approximation for theoretical models of the labor market (though see also Coate and Loury, 1993; Lepage, 2021b), there are many settings in which individuals may have insufficient experience or information to form statistically accurate beliefs. This includes any situation that is novel for the decision-maker, or even familiar settings where some features of the setting is novel. For example, even an individual in human resources with a huge amount of experience making hiring decisions and forming subjective beliefs about the association between qualifications and abilities in a particular job will be unlikely to hold statistically accurate beliefs if they move to a different industry or country. In such settings, as in our experiment, decision-makers face a fair amount of uncertainty about the correct decision which may be an important ingredient for implicit discrimination.

7 Conclusion

The economics literature has traditionally dichotomized discrimination into taste-based (Becker, 1957) and (accurate) statistical discrimination (Phelps, 1972; Arrow, 1973). Even though our experimental design rules out these two classical forms of discrimination, we still observe substantial discrimination against women in the hiring task. In particular, we observe evidence of both explicit and implicit belief-based discrimination. Our classification exercise shows that when hiring decisions are revealing about an individual's gender bias (explicit dis-

crimination), we observe one group of employers who discriminate against women and another smaller group who discriminate against men. However, strikingly, both these groups display a substantial gender bias against women in the complex decisions, where attribution to gender information is obscured. For the employers who switch from explicitly discriminating against men to discriminating against women in the complex decision, this reflects implicit discrimination. Taken together, our results are consistent with the idea that a majority of employers hold a statistically inaccurate gender stereotype that women perform worse in a logic task, but that a subset of employers are reticent to make hiring choices that clearly reveal that they hold this stereotype. Further, our results highlight the importance of the choice setting for determining whether and how discrimination will manifest.

7.1 Implications of our results

There are several important implications of the evidence reported in this paper. First, our results demonstrate that discrimination can take very different forms beyond the traditional distinction between taste-based and (accurate) statistical discrimination. This is important because to choose the correct policy instrument to address discrimination in a particular context, it is imperative to understand the root cause of the problem.[38] Otherwise, the treatment may be ineffective or lead to undesired consequences. For example, policy makers that wish to fight discrimination may be tempted to impose rules that address explicit discrimination, such as: "if choosing between an equally qualified man and woman, choose the woman."[39] While this may be effective in some contexts, in most real-world hiring decisions the candidates differ on many dimensions, so that the evaluation of candidates' overall suitability for a position depends on the rather malleable and subjective relevance assigned to their different attributes. In such contexts with highly heterogeneous candidates, an affirmative action policy rule of this type might be ineffective for changing hiring behavior, while creating the illusion that discrimination is being addressed. Further, it may also lead to individuals going to greater lengths to mask or obscure their discriminatory decisions.

Second, the manifestation of discrimination (both its occurrence and its form) depends crucially on the choice setting. In the same pool of participants, we document evidence of aggregate gender bias when candidates are *differently qualified*, but not when one candidate is *more qualified*.[40] This suggests that discrimination is more likely to occur in settings where candidates are heterogeneous on multiple job-relevant attributes (horizontal heterogeneity), and less likely to occur when candidates are heterogeneous on a single dimension (vertical heterogeneity). This has meaningful implications for the design of remedies that involve altering the

[38] Bohren et al. (2022) provide an insightful discussion of the importance of correctly identifying the source of discrimination in order to design an effective policy intervention to address it.

[39] Cunningham and de Quidt (2023) refer to these as *ceteris paribus* rules. We will use a more general terminology and refer to them as "affirmative action rules."

[40] In addition, we also document evidence of gender bias when candidates are *equally qualified*. This indicates that discrimination is also an issue when candidates are highly homogeneous (in terms of their suitability for the job).

architecture of the choice environment. In particular, situations with horizontal heterogeneity can be translated into situations with vertical heterogeneity by means of carefully designed procedures. For example, one potential solution is to require ex ante criteria that specify how to evaluate candidates on different dimensions, and how to aggregate these evaluations into a single score, as is sometimes already the case in university acceptances. This would remove scope for ex-post re-weighting the different attributes (e.g., as discussed in Hodson, Dovidio, and Gaertner, 2002 and Uhlmann and Cohen, 2005).

Third, our results speak to the long history of theories of human behavior that posit a tension between hidden and expressed motives – ranging from Freud to the modern widespread usage of the IAT in social psychology (Greenwald, McGhee, and Schwartz, 1998). One of the key tensions studied in this social psychology literature is the underlying conflict between explicit egalitarian beliefs and implicit racial biases (Hodson et al., 2010). More recently, the IAT has been used as an effective tool for studying the influence of implicit stereotypes in the economics literature. For example, Carlana (2019) shows that the gender stereotypes of teachers can have a substantial impact on the outcomes of their students (increasing the gender gap in math performance, and inducing girls to self-select into less ambitious high school tracks).[41] The IAT aims to assess the strength of associations between concepts (e.g., "female" / "male" and "logic") by measuring response times when participants classify concepts together in a computerized task. Here, we demonstrate a complementary approach to eliciting implicit preferences from actual choice data, as also suggested by Cunningham and de Quidt (2023) who establish its theoretical foundations. Since the efficacy of the IAT in predicting real-world discrimination is still a contentious topic (see, e.g., Oswald et al., 2015 and Kurdi et al., 2019), a behavioral measure provides an instructive complement to the IAT. Since implicit preferences are by their nature difficult to detect, it is useful to have different measurement tools, each of which may be more appropriate for a particular subset of research contexts.[42]

Fourth, it is important to note that there is no intrinsic reason why the psychological phenomena that we study are specific to labor markets. There are a wide array of domains in which individuals may hold (potentially subconscious) gender stereotypes. If societal norms/laws censure/forbid taking actions on the basis of these gender stereotypes, then individuals face a psychological tension between what is socially appropriate and what they would be inclined to do in the absence of societal prescriptions. Implicit discrimination may manifest in any scenario with this set of characteristics. This could have implications for decisions made in a diverse range of contexts outside of the labor market, such as college admissions, housing rental markets, research funding decisions, immigration decisions, and law enforcement deci-

[41]The IAT has also been used to study implicit racial or ethnic bias by, e.g., Rooth (2010), Glover, Pallais, and Pariente (2017), Corno, La Ferrara, and Burns (2019) and Alesina et al. (2018). The results in Alesina et al. (2018) highlight the immense importance of both: (i) knowing how to detect different forms of discrimination, and (ii) tailoring the policy response to the specific form of discrimination. In their study, teachers who are simply made aware of their implicit biases reduce their discriminatory grading behavior.

[42]For example, when designing surveys, using the IAT to measure a respondent's implicit biases may be impractical, but adding a few carefully designed (hypothetical) choice questions that vary in how strongly they reveal the decision maker's motives may well be feasible.

sions.

Finally, it is worth noting that while we rule out taste-based discrimination by design in this paper, and study the tension that may arise between stereotypes (beliefs) and societal norms, implicit discrimination may also result from a tension between preferences and societal norms.[43] For example, if an individual simply dislikes individuals from a particular group, but societal norms prescribe non-discrimination, the individual may (consciously) discriminate when their discrimination is obscured, but not in scenarios where it will be revealed to others. This is subtly different from the type of belief-based discrimination that we study but also important.

7.2 Limitations and future research

One important caveat to our results is that the degree of implicit discrimination that we detect is likely to be underestimated in relation to its occurrence in natural settings in the general population. There are several reasons for this. First, our population is comprised of young and highly educated students who are likely to hold less gender-stereotyped beliefs than the general population.[44] Second, the hiring decisions that participants make in our experiment are anonymous. Therefore, the role of social image concerns is substantially dampened. Since implicit discrimination involves a tension between an underlying preference and the signal that one's actions send (to oneself and others), the dampening of social image concerns is likely to yield a shift towards more explicit discrimination and less implicit discrimination. In many real-world contexts, decisions are not anonymous, and one would expect that the increased role of social image would lead to a shift away from explicit discrimination. At the same time, real-world hiring decisions are typically complex, especially among the "finalists" in a hiring process. In such scenarios, implicit discrimination is likely to play a larger role. Together, these considerations point towards the worrying conclusion that if we are able to detect implicit discrimination in the stark, anonymous environment of our experiment, it is likely to be substantially more prevalent in real world contexts.

A key question to address in future research, therefore, is: What are the contextual and institutional factors that are likely to generate implicit discrimination? As implicit discrimination can result from a conflict between what an individual would like to do (preferences), and

[43] It is also worth noting that the line between preference-based and belief-based forms of discrimination can become fuzzy if one allows for motivated beliefs or belief-based utility. An individual who gains utility from believing that their own group is better than another group may form motivated stereotypes in order to increase their belief-based utility. Under this framework, the distinction between belief-based and taste-based discrimination becomes less clear-cut.

[44] A large fraction of the participants in our experiment attend a technical university, implying that they interact regularly with male and female classmates that are selected to be above-average in terms of their quantitative abilities. This may serve to ameliorate gender stereotypes they previously held.

what is socially acceptable behavior (norms),[45] it follows that it is more likely to be observed in hiring scenarios with the following characteristics. Scenarios where: (i) preferentially hiring a candidate from a particular group is socially stigmatized, (ii) many individuals in the population of decision makers hold stereotypes (or tastes) that favor this group, (iii) the job candidates are (horizontally) heterogeneous, or the expertise and attributes required for the job are more opaque (i.e., the "revealingness" of the hiring decisions about biases is low).[46]

7.3 Lessons for policy design

One important lesson from the recent discrimination literature is that it is imperative that policy interventions are tailored to address the source of the problem. In the case of implicit discrimination, the design of policy interventions depends critically on whether individuals are masking their discriminatory preferences from themselves (self-image) or others (social image) – i.e., whether they are really aware of their bias or not. In situations where individuals are unaware of their own bias, it may be sufficient to inform these individuals about the bias present in their own or other individuals' past decision-making. Alesina et al. (2018) demonstrate that this can be effective in de-biasing teachers with implicit discriminatory preferences. If, instead, individuals are fully aware of their bias and are hiding their preferences from others, the policy prescription is very different. Here, carefully designed procedures, such as requiring clear and transparent ex-ante decision rules that leave little wiggle room for evaluators might be more effective (see, e.g., Uhlmann and Cohen, 2005). In cases where inaccurate gender-biased beliefs or stereotypes are at the heart of discrimination, as presented here, confronting these beliefs with information can also be an effective approach. This solution is discussed by Bohren et al. (2022). Bordalo et al. (2016) argue that stereotypes are typically based on a "kernel of truth".[47] If one can demonstrate in a particular context that there are no relevant statistical differences between two groups, this may induce a re-evaluation of the stereotype.

[45] In situations with these characteristics, the motive to discriminate explicitly is reduced by social stigma. For example, Barr, Lane, and Nosenzo (2018) provide evidence that discrimination is reduced when it is perceived to be more socially inappropriate (although, they focus on taste-based discrimination). We argue here that depending on the context, these underlying preferences may instead manifest as implicit instead of explicit discrimination.

[46] Another important factor that we do not analyze in our study is the role of the employer's gender in generating explicit and implicit discrimination. While we think that this would be an interesting dimension to examine, it would require having precise ex-ante hypotheses about the ways that employer gender could interact with the mechanisms we study. Further, it is not clear that gender differences should be expected at all. For example, psychological studies on stereotypes in the workplace commonly describe a "lack of differences between women and men as evaluators" (Heilman, 2012, p. 129). Studying gender differences in employer behavior in our experiment would involve collecting a much larger sample to facilitate testing gender-specific hypotheses. We, therefore, simply note the following suggestive finding on the matter, leaving a thorough investigation to future research: When having to decide between a male and a female candidate (which is the case in 7 out of the 9 hiring decisions, see Table 1), male employers choose the male candidate 56.2% of the time, whereas female employers choose the male candidate 50.7% of the time.

[47] However, it is important to note that the "kernel of truth" may be the result of endogenous processes in society that make stereotypes self-fulfilling. For example, Chauvin (2018) demonstrates that in a society where individuals are prone to exhibit the Fundamental Attribution Error, they underestimate the role played by differing circumstances on the outcomes of different groups, and therefore form biased beliefs about underlying characteristics of these groups.

However, discriminatory beliefs can be sticky even in the presence of informative signals that contradict them (Reuben et al., 2014). This may especially be the case when a motivation exists to maintain false beliefs against incoming data (as for favorable in-group beliefs, demonstrated by Cacault and Grieder, 2019). Such motivated tastes over beliefs are harder to combat – doing so requires influencing the formation of preferences, which is a complex process taking place over a long period of time and not easy to influence. Lai et al. (2016) show that brief interventions like presenting counter-stereotypical examples are unlikely to have long-lasting impacts on implicit bias. Further, Dovidio et al. (2016) discuss how many well-intentioned interventions aimed at reducing intergroup bias may backfire.

Interestingly, our results imply that hiring procedures that force joint rather than separate evaluation of candidates, as suggested by the lab experiments of Bohnet et al. (2015), are not a panacea when performance signals are less straightforward to interpret (i.e., there is not a clear and simple correspondence between qualifications and the job being hired for) and do not allow one to unambiguously rank one candidate over the other. In line with their results, though, we find no gender bias in joint evaluations of female-male candidate pairs where ranking by qualification is simple (in our case, one certificate signal vs. no certificate signal).

Thus, together with the contemporary discrimination literature, this paper highlights that in order to find effective remedies to combat discrimination, it is crucial to have a fine-grained and accurate understanding of the underlying causes of discrimination and to be able to detect the different manifestations that discriminatory preferences can take in different contexts. Implicit discrimination is a particularly problematic form of discrimination because: (i) by its nature, it is even more difficult to identify, and therefore regulate, than explicit forms of discrimination are, and (ii) it is likely to materialize in precisely the contexts where explicit discrimination has already been acknowledged to be unethical and is therefore highly stigmatized. The paper also demonstrates a central role of beliefs in the formation of discriminatory behavior. Future work in this area might investigate the relative importance of self-image and social-image in generating implicit discrimination, and systematically study the contextual and institutional factors that exacerbate and alleviate it. Lessons learned from these exercises would be invaluable for designing effective policy tools that are able to treat the underlying problem, as opposed to just treating the symptoms and allowing discrimination to simply manifest in a different form.

References

Agresti, A. (2013). *Categorical data analysis* (3 ed.). John Wiley & Sons.

Alesina, A., M. Carlana, E. La Ferrara, and P. Pinotti (2018). Revealing stereotypes: Evidence from immigrants in schools. *NBER Working Paper 25333*.

Arrow, K. J. (1973). The theory of discrimination. In O. Ashenfelter and A. Rees (Eds.), *Discrimination in Labor Markets*. Princeton University Press.

Ayres, I. and P. Siegelman (1995). Race and gender discrimination in bargaining for a new car. *American Economic Review 85*(3), 304–321.

Banaji, M. R. and A. G. Greenwald (1995). Implicit gender stereotyping in judgments of fame. *Journal of Personality and Social Psychology 68*(2), 181.

Barr, A., T. Lane, and D. Nosenzo (2018). On the social inappropriateness of discrimination. *Journal of Public Economics 164*, 153–164.

Becker, G. S. (1957). *The Economics of Discrimination*. Chicago: University of Chicago Press.

Bertrand, M., D. Chugh, and S. Mullainathan (2005). Implicit discrimination. *American Economic Review 95*(2), 94–98.

Bertrand, M. and E. Duflo (2017). Field experiments on discrimination. In A. Banerjee and E. Duflo (Eds.), *Handbook of Field Experiments*. North Holland.

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

Biernat, M. and D. Kobrynowicz (1997). Gender-and race-based standards of competence: Lower minimum standards but higher ability standards for devalued groups. *Journal of Personality and Social Psychology 72*(3), 544.

Blau, F. D. and L. M. Kahn (2017). The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature 55*(3), 789–865.

Bohnet, I., A. Van Geen, and M. Bazerman (2015). When performance trumps gender bias: Joint vs. separate evaluation. *Management Science 62*(5), 1225–1234.

Bohren, A., K. Haggag, A. Imas, and D. G. Pope (2022). Inaccurate statistical discrimination. *Review of Economics and Statistics (forthcoming)*.

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

Bohren, J. A., P. Hull, and A. Imas (2022). Systemic discrimination: Theory and measurement. *Mimeo*.

Bordalo, P., K. Coffman, N. Gennaioli, and A. Shleifer (2016). Stereotypes. *Quarterly Journal of Economics 131*(4), 1753–1794.

Bordalo, P., K. Coffman, N. Gennaioli, and A. Shleifer (2019). Beliefs about gender. *American Economic Review 109*(3), 739–73.

Bowles, H. R., L. Babcock, and L. 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.

Brief, A. P., J. Dietz, R. R. Cohen, S. D. Pugh, and J. B. Vaslow (2000). Just doing business: Modern racism and obedience to authority as explanations for employment discrimination. *Organizational behavior and human decision processes 81*(1), 72–97.

Cacault, M. P. and M. Grieder (2019). How group identification distorts beliefs. *Journal of Economic Behavior & Organization 164*, 63 – 76.

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

Carlana, M. (2019). Implicit stereotypes: Evidence from teachers' gender bias. *Quarterly Journal of Economics 134*(3), 1163–1224.

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

Chauvin, K. P. (2018). A misattribution theory of discrimination. *Mimeo*.

Chen, D. L., M. Schonger, and C. Wickens (2016). oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance 9*, 88–97.

Coate, S. and G. C. Loury (1993). Will affirmative-action policies eliminate negative stereotypes? *American Economic Review 83*, 1220–1240.

Coffman, K., M. Collis, and L. Kulkarni (2021). Stereotypes and belief updating. *Mimeo*.

Coffman, K. B., P. U. Araya, and B. Zafar (2021). A (dynamic) investigation of stereotypes, belief-updating, and behavior. *NBER Working Paper 29382*.

Coffman, K. B., C. L. Exley, and M. Niederle (2021). The role of beliefs in driving gender discrimination. *Management Science 67*(6), 3551–3569.

Coffman, K. B., C. B. Flikkema, and O. Shurchkov (2021). Gender stereotypes in deliberation and team decisions. *Games and Economic Behavior (forthcoming)*.

Corno, L., E. La Ferrara, and J. Burns (2019). Interaction, stereotypes and performance: Evidence from South Africa. *IFS Working Papers*.

Cunningham, T. and J. de Quidt (2023). Implicit preferences inferred from choice. Mimeo.

De Franca, D. X. and M. B. Monteiro (2013). Social norms and the expression of prejudice: The development of aversive racism in childhood. *European Journal of Social Psychology 43*(4), 263–271.

Dovidio, J. F. and S. L. Gaertner (2000). Aversive racism and selection decisions: 1989 and 1999. *Psychological Science 11*(4), 315–319.

Dovidio, J. F., S. L. Gaertner, and A. R. Pearson (2017). Aversive racism and contemporary bias. In C. Sibley and F. Barlow (Eds.), *The Cambridge Handbook of the Psychology of Prejudice*, pp. 267–294. Cambridge University Press.

Dovidio, J. F., S. L. Gaertner, E. G. Ufkes, T. Saguy, and A. R. Pearson (2016). Included but invisible? Subtle bias, common identity, and the darker side of "we". *Social Issues and Policy Review 10*(1), 6–46.

Dustan, A., K. Koutout, and G. Leo (2022, August). Second-order beliefs and gender. *Journal of Economic Behavior & Organization 200*, 752–781.

Erkal, N., L. Gangadharan, and B. H. Koh (2021). Gender biases in performance evaluation: The role of beliefs versus outcomes. *Available at SSRN 3979701*.

Esponda, I., R. Oprea, and S. Yuksel (2022). Discrimination without reason: Biases in statistical discrimination. *Mimeo*.

Esponda, I., R. Oprea, and S. Yuksel (2023). Seeing what is representative. *The Quarterly Journal of Economics*, qjad020.

Eyting, M. (2022). Why do we discriminate? the role of motivated reasoning. *SAFE Working Paper No. 356*.

Fang, H. and A. Moro (2011). Chapter 5 - theories of statistical discrimination and affirmative action: A survey. Volume 1 of *Handbook of Social Economics*, pp. 133–200. North-Holland.

Gaertner, S. L. and J. F. Dovidio (2000). The aversive form of racism. In C. Stangor (Ed.), *Stereotypes and prejudice: Essential readings*, pp. 289–304. Psychology Press.

Glick, P., C. Zion, and C. Nelson (1988). What mediates sex discrimination in hiring decisions? *Journal of Personality and Social Psychology 55*(2), 178.

Glover, D., A. Pallais, and W. Pariente (2017). Discrimination as a self-fulfilling prophecy: Evidence from french grocery stores. *Quarterly Journal of Economics 132*(3), 1219–1260.

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

Greenwald, A. G., D. E. McGhee, and J. L. Schwartz (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology 74*(6), 1464.

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

Heilman, M. E. (2012, January). Gender stereotypes and workplace bias. *Research in Organizational Behavior 32*, 113–135.

Hengel, E. (2022). Publishing while female. Are women held to higher standards? Evidence from peer review. *Mimeo*.

Hilton, J. L. and W. Von Hippel (1996). Stereotypes. *Annual Review of Psychology 47*(1), 237–271.

Hodson, G., J. F. Dovidio, and S. L. Gaertner (2002). Processes in racial discrimination: Differential weighting of conflicting information. *Personality and Social Psychology Bulletin 28*(4), 460–471.

Hodson, G., J. F. Dovidio, and S. L. Gaertner (2010). The aversive form of racism. In J. L. Chin (Ed.), *Race and ethnicity in psychology. The psychology of prejudice and discrimination: Racism in America*, Volume 1, pp. 119–135. Praeger Publishers.

Hodson, G., H. Hooper, J. Dovidio, and S. Gaertner (2005). Aversive racism in britain: Legal decisions and the use of inadmissible evidence. *European Journal of Social Psychology 35*(4), 437–448.

Hübert, R. and A. T. Little (2023). A behavioural theory of discrimination in policing. *The Economic Journal*, uead043.

Isaksson, S. (2018). It takes two: Gender differences in group work. *Mimeo*.

Jowell, R. and P. Prescott-Clarke (1970). Racial discrimination and white-collar workers in britain. *Race 11*(4), 397–417.

Judd, C. M. and B. Park (1993). Definition and assessment of accuracy in social stereotypes. *Psychological Review 100*(1), 109.

Kleinpenning, G. and L. Hagendoorn (1993). Forms of racism and the cumulative dimension of ethnic attitudes. *Social Psychology Quarterly*, 21–36.

Kovel, J. (1970). *White racism: A psychohistory*. New York: Pantheon.

Kübler, D., J. Schmid, and R. Stüber (2018). Gender discrimination in hiring across occupations: a nationally-representative vignette study. *Labour Economics 55*, 215–229.

Kurdi, B., A. E. Seitchik, J. R. Axt, T. J. Carroll, A. Karapetyan, N. Kaushik, D. Tomezsko, A. G. Greenwald, and M. R. Banaji (2019). Relationship between the implicit association test and intergroup behavior: A meta-analysis. *American Psychologist 74*(5), 569.

Lai, C. K., A. L. Skinner, E. Cooley, S. Murrar, M. Brauer, T. Devos, J. Calanchini, Y. J. Xiao, C. Pedram, C. K. Marshburn, S. Simon, J. C. Blanchar, J. A. Joy-Gaba, J. Conway, L. Redford, R. A. Klein, G. Roussos, F. M. H. Schellhaas, M. Burns, X. Hu, M. C. McLean, J. R. Axt, S. Asgari, K. Schmidt, R. Rubinstein, M. Marini, S. Rubichi, J.-E. L. Shin, and B. A. Nosek (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General 145*(8), 1001–1016.

Lane, T. (2016). Discrimination in the laboratory: A meta-analysis of economics experiments. *European Economic Review 90*, 375–402.

Lepage, L.-P. (2021a). Bias formation and hiring discrimination. *Mimeo*.

Lepage, L.-P. (2021b). Endogenous learning, persistent employer biases, and discrimination. *Mimeo*.

McIntyre, S., D. J. Moberg, and B. Z. Posner (1980). Preferential treatment in preselection decisions according to sex and race. *Academy of Management Journal 23*(4), 738–749.

Mengel, F. and P. Campos-Mercade (2023). Non-bayesian statistical discrimination. *Management Science (forthcoming)*.

Milkman, K. L., M. Akinola, and D. 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), 1678.

Nadler, J. T., M. R. Lowery, J. Grebinoski, and R. G. Jones (2014). Aversive discrimination in employment interviews: Reducing effects of sexual orientation bias with accountability. *Psychology of Sexual Orientation and Gender Diversity 1*(4), 480.

Neumark, D. (2018). Experimental research on labor market discrimination. *Journal of Economic Literature 56*(3), 799–866.

Neumark, D., R. J. Bank, and K. D. Van Nort (1996). Sex discrimination in restaurant hiring: An audit study. *Quarterly Journal of Economics 111*(3), 915–941.

Newman, J. M. (1978). Discrimination in recruitment: An empirical analysis. *Industrial and Labor Relations Review 32*(1), 15–23.

Norton, M. I., J. A. Vandello, and J. M. Darley (2004). Casuistry and social category bias. *Journal of Personality and Social Psychology 87*(6), 817.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science 349*(6251), aac4716.

Oswald, F. L., G. Mitchell, H. Blanton, J. Jaccard, and P. E. Tetlock (2015). Using the IAT to predict ethnic and racial discrimination: Small effect sizes of unknown societal significance. *Journal of Personality and Social Psychology 108*(4), 562–571.

Pearson, A. R., J. F. Dovidio, and S. L. Gaertner (2009). The nature of contemporary prejudice: Insights from aversive racism. *Social and Personality Psychology Compass 3*(3), 314–338.

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

Rackstraw, E. (2022). Bias-motivated updating in the labor market. *Available at SSRN 4278076*.

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

Riach, P. A. and J. Rich (1987). Testing for sexual discrimination in the labour market. *Australian Economic Papers 26*(49), 165–178.

Riach, P. A. and J. Rich (2002). Field experiments of discrimination in the market place. *Economic Journal 112*(483), F480–518.

Rooth, D.-O. (2010). Automatic associations and discrimination in hiring: Real world evidence. *Labour Economics 17*(3), 523–534.

Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys* (3 ed.). John Wiley & Sons.

Sarsons, H. (2019). Interpreting signals in the labor market: Evidence from medical referrals. *Working Paper*.

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

Saucier, D. A., C. T. Miller, and N. Doucet (2005). Differences in helping whites and blacks: A meta-analysis. *Personality and Social Psychology Review 9*(1), 2–16.

Savage, Jr., R. P. (1994). The paradox of nontransitive dice. *The American Mathematical Monthly 101*(5), 429–436.

Small, D. A., M. Gelfand, L. Babcock, and H. Gettman (2007). Who goes to the bargaining table? The influence of gender and framing on the initiation of negotiation. *Journal of Personality and Social Psychology 93*(4), 600.

Snyder, M. L., R. E. Kleck, A. Strenta, and S. J. Mentzer (1979). Avoidance of the handicapped: an attributional ambiguity analysis. *Journal of Personality and Social Psychology 37*(12), 2297.

Son Hing, L. S., G. A. Chung-Yan, L. K. Hamilton, and M. P. Zanna (2008). A two-dimensional model that employs explicit and implicit attitudes to characterize prejudice. *Journal of Personality and Social Psychology 94*(6), 971.

Tracy, S. J. and K. D. Rivera (2010). Endorsing equity and applauding stay-at-home moms: How male voices on work-life reveal aversive sexism and flickers of transformation. *Management Communication Quarterly 24*(1), 3–43.

Tversky, A. and D. Kahneman (1981). The framing of decisions and the psychology of choice. *Science 211*, 453–458.

Uhlmann, E. L. and G. L. Cohen (2005). Constructed criteria: Redefining merit to justify discrimination. *Psychological Science 16*(6), 474–480.

van Buuren, S. (2018). *Flexible Imputation of Missing Data* (2 ed.). New York: Chapman and Hall/CRC.

Wojcieszak, M. (2015). Aversive racism in spain-testing the theory. *International Journal of Public Opinion Research 27*(1), 22–45.

Yinger, J. (1986). Measuring racial discrimination with fair housing audits: Caught in the act. *American Economic Review 76*(5), 881–893.

Yu, C. W., Y. J. Zhang, and S. X. Zuo (2021). Multiple switching and data quality in the multiple price list. *Review of Economics and Statistics 103*(1), 135–150.

ONLINE APPENDICES

"Explicit and Implicit Belief-Based Gender Discrimination: A Hiring Experiment"

Kai Barron, Ruth Ditlmann, Stefan Gehrig, Sebastian Schweighofer-Kodritsch

A The Job Candidate Assessment: Tasks and Procedure

A.1 The word task (word certificate)

Participants solved three word search puzzles.[48] They had 90 seconds to work on each puzzle. Each of the puzzles contained 10 hidden words, and participants were presented with 30 possible answers. Participants selected answers they thought were correct. For each correct answer, participants gained 0.40€. For each wrong answer, they lost 0.40€ (we restricted the total payoff in this task to be non-negative). On average, participants earned 5.03€ (SD: 1.80€) in this task. Performance was measured as the total number of selected correct response options minus the number of selected incorrect response options.

A.2 The knowledge task (knowledge certificate)

This task consisted of 30 general knowledge questions, for which four minutes were available. Questions were selected from several categories (geography, environmental sciences, pop culture, arts, literature and history) but were presented in an arbitrary order. Four response options were presented for each question, of which only one was correct. Each correct answer was worth 0.60€. On average, participants earned 5.08€ (SD: 2.14€) in this task. Performance was measured as total number of questions answered correctly.

A.3 The logic task (job task)

Participants solved matrix reasoning exercises of the type that are commonly used in general intelligence tests. Each of the ten questions consisted of a 3-by-3 matrix in which one cell was empty. Matrices had to be completed by choosing one of the six response options.[49] Participants were given five minutes to work on this task. They earned 1.30€ for each matrix problem they solved correctly. On average, they earned 5.22€ (SD: 1.87€) in this task. Performance was measured as total number of matrix exercises solved correctly.

A.4 Procedure in the Job Candidate Assessment

The order of the tasks was held constant for all participants. After the completion of all tasks, an incentivized belief elicitation and questions on demographics followed. For each of the tasks mentioned above, participants were asked how often they believed a randomly drawn male would perform better than a randomly drawn female.[50] They were also asked what

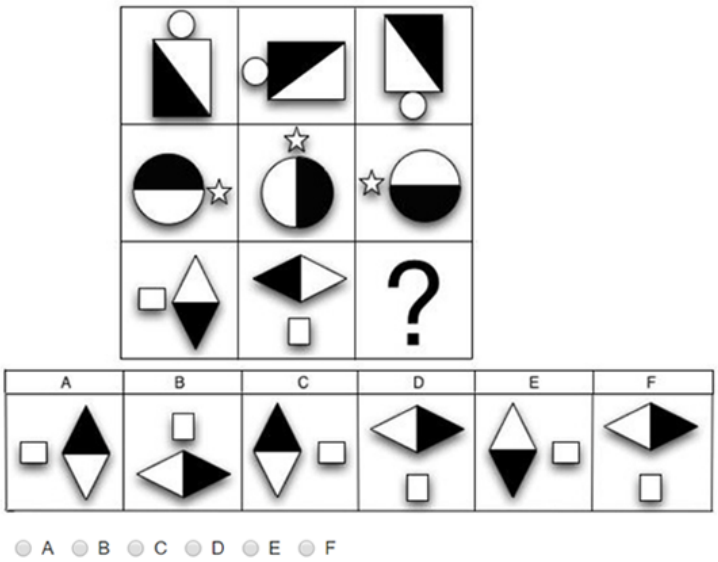
[48] Each puzzle had a theme: animals, countries or fruit.

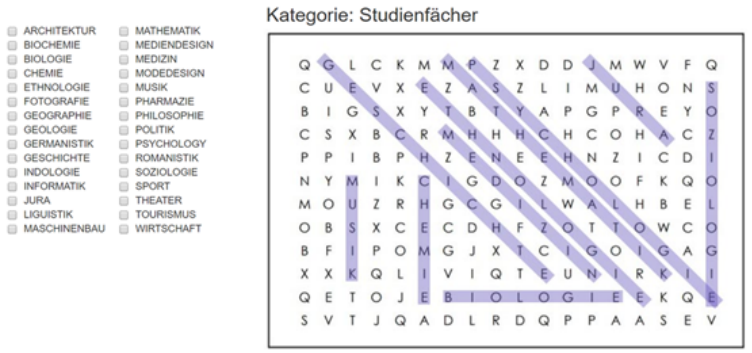
[49] Matrix exercises were taken from the online resources of the ICAR project. See: `https://icar-project.com/projects/icar-project`

[50] More specifically, we asked participants to think about taking 100 draws of a pair of participants, each containing a randomly drawn male and a randomly drawn female from their session. They were asked to indicate how often they believed that the randomly drawn male performed better than the randomly drawn female in the respective task (with ties broken randomly).

they thought all other people in the experiment on average responded to the previous belief elicitation. Participants' beliefs were incentivized by means of a quadratic scoring rule.

Figure 4: Examples of the logic task (upper panel), the word task (middle panel) and the knowledge task (lower panel). These examples and their solutions were shown to participants in the JOB CANDIDATE ASSESSMENT as part of the instructions before they worked on the problems they were scored for.





B Relation to the Psychology Literature – Further Details

In social psychology, there exists a body of work that is closely related to the idea of implicit discrimination as conceptualized in our paper. This literature in social psychology is associated with the "aversive racism" framework. The term aversive racism was coined by the psychoanalyst Joel Kovel (1970) and systematically investigated in the 1980s by John Dovidio and Samuel Gaertner. It explains how people who want to seem fair and equal, but hold negative feelings towards a particular race deep down, end up discriminating against Black people in subtle ways that are not easily recognized as racial discrimination. This implicit discrimination allows them to express their racial bias without being recognized as racist. Aversive racists do not experience hatred towards Black people but anxiety, fear, or discomfort. "While they find Blacks aversive, they find any suggestion that they might be prejudiced aversive as well" according to Pearson, Dovidio, and Gaertner (2009; p. 271). Because they try to avoid appearing racist, they won't discriminate and might even overcompensate when decisions are obviously based on race. But when the situation is unclear, and they can find a non-racial reason for their biased behavior, aversive racists will end up discriminating against Black individuals (Pearson et al., 2009).

Aversive racism was originally discovered in experiments on intergroup helping. Most experiments on intergroup helping found that White participants only helped Black actors when not doing so would have made them appear racially biased (Saucier, Miller, and Doucet, 2005). Similarly, in a different experiment non-disabled participants chose to sit further away from a disabled vs. non-disabled confederate when they were given a reason to justify this choice by the experimenter (Snyder et al., 1979).

The first experiment on intergroup selection was conducted in 1988-89 and replicated in 1998-1999. In this experiment, Dovidio and Gaertner (2000) presented White undergraduate students (N=194) with applications of fictitious applicants for the role of peer counselor. The participants then had to indicate on rating scales how qualified the candidate was, and if they would recommend hiring him. The fictitious job candidates were either Black or White as suggested by the activities listed on their resumes (e.g., membership in Black student union), and their applications had either no flaws, minor flaws, or serious flaws indicated by their response to a hypothetical job scenario. In support of aversive racism, job candidates preferred the White candidate in the condition with minor flaws, but not in the other two conditions, where such a preference would unambiguously indicate racial bias (Dovidio and Gaertner, 2000).

This selection experiment has since been replicated in England (Hodson, Hooper, Dovidio, and Gaertner, 2005) and Spain (Wojcieszak, 2015) and with homosexual applicants (Nadler, Lowery, Grebinoski, and Jones, 2014). Other implications of the aversive racism framework have been tested in different countries, such as Canada (Son Hing, Chung-Yan, Hamilton, and Zanna, 2008), the Netherlands (Kleinpenning and Hagendoorn, 1993), and Portugal (De Franca and Monteiro, 2013).

To date, there has been no exact replication of the selection paradigm in psychology with female vs. male candidates, although the concept of aversive sexism is sometimes used to explain differential outcomes for women in the workplace. For example, a discursive study of male business executives found that consistent with aversive sexism, they espouse gender equality at work while subtly preferring traditional gender roles at home (Tracy and Rivera, 2010).

A proposed mechanism for the observed aversive racism effects is a shift in how relevant criteria for the job are interpreted. Indeed, several experiments show that if a candidate belongs to a minority group (rather than a majority group), evaluators prioritize the criteria where the candidate has flaws (Hodson et al., 2002; see also Brief, Dietz, Cohen, Pugh, and Vaslow, 2000, for employment bias against Blacks; Rooth, 2010, for hiring biases against Muslims). This effect was also found in employment and college application decisions with male and female candidates in laboratory experiments (Norton et al., 2004; Uhlmann and Cohen, 2005).

Later studies on aversive racism replaced the behavioral selection paradigm, where aversive racism is inferred from behavior, with a strategy for measuring aversive racism directly. Through measuring implicit race-bias with an IAT and explicit egalitarian values, researchers categorize participants into four groups: Those who display both implicit and explicit bias (modern racists), those who do not show bias on either measure (truly low prejudiced), those who show bias on the explicit but not implicit measures (principled conservatives) and finally, the aversive racist group that shows racial bias on implicit but not explicit measures (Son Hing et al., 2008).

The primary contribution of the present research relative to this body of work is to introduce the concept of aversive sexism to the economics literature, using the terminology, theories, and methods of behavioral economics. This is important because re-interpreting these ideas through the lens of behavioral economics is non-trivial. In doing this, we extend some of the ideas, for example, by drawing on the behavioral identification approach proposed by Cunningham and de Quidt (2023) using non-transitive cycles and implementing this in our empirical analysis.

In addition to incorporating existing ideas from psychology into behavioral economics, our study also contributes directly to the social psychology literature in several important ways: First, we are the first to systematically study the full aversive sexism framework, including ambiguous (complex) and unambiguous (gender) decisions, and to identify both explicit and implicit discrimination in the same experiment. Second, our study provides a form of stress test of these ideas from social psychology, since we test them: (i) using the methods of experimental economics, (ii) in a different national context (Germany), and (iii) concerning a different group (women). Given the recent replication crisis (Open Science Collaboration, 2015), it is reassuring to see that we find evidence that is consistent with the seminal work on aversive racism in our very different context using different methods. Third, our experiment goes beyond the research in Social Psychology by giving incentives for accurate choices. Doing so, allows us to demonstrate that social image concerns trump accuracy incentives. When thinking about the

research in social psychology that is unincentivized, the omission of accuracy incentives may have implications for generalizing from the laboratory to field settings (and may also be one potential explanation for why explicit discrimination is sometimes not detected in experiments where there is an absence of incentives for accuracy). Our findings help to alleviate such concerns. While the results from our study should also be interpreted with the relevant caveats associated with any laboratory study, our findings offer valuable insight by demonstrating the persistent nature of implicit bias even when incentives for accuracy are introduced.

C Identification – Gender Bias and Discrimination

Table 3: Identification of gender bias and discrimination.

	(f,K)	(m,W)
(m,K)	σ_K	σ_m
(f,W)	σ_f	σ_W

We formally illustrate the relationship between aggregate level gender bias and individual level discrimination for the complex decisions using Table 3. The row indicates the preferred candidate in Com_{fW}, and the column indicates the preferred candidate in Com_{fK}. The cells indicate the proportions of employers for all four possible preference combinations, i.e., the joint distribution over preference types (all four entries sum to one). The two on-diagonal proportions σ_K and σ_W are those employers who would consistently choose one qualification over the other, regardless of gender, and hence do not discriminate. The two off-diagonal proportions σ_f and σ_m are those employers who would consistently choose one gender over the other, regardless of qualification, and hence do discriminate (σ_f against men and σ_m against women).

Let x be the fraction of employers who would choose the K-candidate in the "row decision" (male) and similarly let $(1-x')$ denote the fraction who would choose the K-candidate in the "column decision" (female). This allows us to define $\overline{x} := (x+x')/2$ as the average rate at which men are hired across the two decisions/treatments. Given that treatment assignment is random, by observing these fractions, we obtain unbiased estimates of the marginals $x = \sigma_K + \sigma_m$ and $(1-x') = \sigma_K + \sigma_f$. Differencing out σ_K yields,

$$b := 2 \cdot (\overline{x} - 0.5) \equiv x - (1-x') = \sigma_m - \sigma_f \tag{1}$$

which is our measure of gender bias, and which we can directly observe in our data. If no employers were to discriminate, i.e., if $\sigma_m = \sigma_f = 0$, then we would observe $b = 0$. Any significant gender bias $b \neq 0$ therefore identifies the presence of relative discrimination at the aggregate level: $b > 0$ implies $\sigma_m > \sigma_f$ and hence that there are more employers discriminating against women than there are employers discriminating against men, and vice versa if $b < 0$.[51] Furthermore, the fact that we always have $\sigma_f \geq 0$ implies $\sigma_m \geq b$, which means that the observed aggregate gender bias b provides us with a lower bound on the proportion of employers that discriminate against women. In other words, b provides an estimate of relative discrimination against women and thereby provides a lower bound on the amount of absolute discrimination against women. Similarly, using the observation that $\sigma_m \geq 0$, we obtain $\sigma_f \geq -b$. (Note that one such lower bound is positive if and only if the other is negative.)

[51] While gender bias implies discrimination, the converse is not true. If there is an equal amount of discrimination against men and women, then $b = 0$. In general, $\sigma_m = \sigma_f$ implies $x = (1-x')$. As an extreme example, if half of all employers prefer men irrespective of qualification, and half do so for women, then we will observe no gender bias even though every employer discriminates ($\sigma_m = \sigma_f = 1/2$).

Of course, x and x' immediately yield also upper bounds on absolute discrimination against women, namely $\sigma_m \leq \min\{x, x'\}$ and $\sigma_f \leq \min\{1-x, 1-x'\} = 1-\max\{x, x'\}$.

D Robustness Exercises

Our identification of discrimination does not rely on imposing assumptions on subjective beliefs. However, it is important to acknowledge that in the hiring decisions, the employers do not know the statistical relationship between gender or the certificates and job performance. This implies that they need to form their own subjective beliefs regarding these statistical relationships. While this is also the case in many real-world settings where individuals do not know the statistical relationship between variables, one may be concerned that this subjectivity introduces noise in employers' decision-making. A second concern is that while the explicit discrimination type classification we use above is arguably natural, it is not the only option for conducting the classification. To alleviate these concerns, we now present various robustness exercises that provide support for our main finding, namely identifying the presence of implicit discrimination against women.

Ultimate hiring propensities. In our main analysis, we focus on the initial choices made in the complex decisions. These decisions were the first ones employers made, and they were unaware that they would subsequently have the opportunity to sell their choices. Under the assumption that no one is entirely indifferent, we can interpret this initial binary choice as a strict preference, which simplifies the examination of gender bias and discrimination in these between-subjects decisions. If all employers were "highly uncertain" (i.e., close to indifferent in this decision), then it is plausible that our finding would be more spurious than robust.

Reassuringly, however, the gender bias in the ultimate hiring choices in the complex decisions is very similar to that in the initial choices.[52] Specifically, when we consider the ultimate decisions, the propensity to hire the male candidate remains nearly unchanged for the two explicit discrimination types we are predominantly interested in – those who discriminate explicitly against men and those who discriminate explicitly against women: In decision Com_{fW}, the two types' ultimate propensities of hiring the male candidate are 59.4% and 61.7%, respectively, and in decision Com_{fK}, these are 66.7% and 68.8%.

Therefore, the gender bias in the complex decisions remains similar for these two groups of explicit gender discriminators, at 26.0% and 30.4%, respectively. The former is significantly larger than zero, therefore establishing implicit discrimination against women even under this further restriction ($z = 1.45$, $p = 0.074$).[53] The reason is simply that few of these explicit gender discriminating employers sell their initial choice (3 out of 31, and 5 out of 62, respec-

[52]This can also be seen in Figures 10 and 11 in Appendix G. Figure 10 provides a visual illustration of the distribution of explicit discrimination types and shows how this distribution relates to both initial and ultimate hiring propensities in the complex decisions.

[53]The measure of ultimate male bias constitutes a lower bound on the fraction of employers with a male bias in the complex decisions. Besides those "insisting" on the male candidate in both decisions, this includes also employers that insist on the man in one decision while selling their choice in the other. Because the measure counts the latter essentially as "half" discriminating against women only, and the gender bias measure also subtracts the mirror image of employers with a female bias, we still obtain a lower bound on discrimination against women, but it is not necessarily true anymore that the smaller male hiring propensity between the two decisions yields an upper bound.

tively, for the two types). This is in contrast to the explicit non-discriminators, for whom the small bias against men observed in the initial decisions is non-robust: In their ultimate hiring choices, it reverses to a very small, non-significant bias of 2.9% against women (20 out of 71 sell their initial choice).

Alternative explicit discrimination type classification. To check the robustness of our main finding, detecting implicit discrimination, we consider two alternative approaches to classifying individuals into explicit discrimination types. In our main classification, the group of employers that we define as discriminating explicitly against men includes those employers who sell one choice (but not both). Therefore, as an alternative classification, we consider employers that do not sell either of the two choices. This results in a smaller group of employers—7.6% and 9.0% display this stricter explicit preference against men in the two treatments. However, under this stricter classification exercise, we find that the bias in favor of males in the complex decisions turns out to be even higher. In Com_{fW}, 77.8% choose the male candidate, while 63.6% do so in Com_{fK}. While this stricter classification exercise implies that the analysis is conducted on a smaller sample of employers, it is comforting that the results remain strongly in line with our main finding, still showing significant implicit discrimination against women ($z = 1.84$, $p = 0.033$).

Another potential concern is that employers may strive for gender balance across the hiring decisions that they make. For example, after hiring a man in the first decision (the complex decision), an employer may aim to maintain a gender balance in their "portfolio" by then hiring a woman in the next decision, namely the first gender decision. If this employer were then to indicate indifference in the second gender decision by selling her choice, this could look like implicit discrimination. To examine whether our main result is driven by employers with such a balancing heuristic, we exploit the order in which gender decisions are made. To do this, we first evaluate the male bias in the complex decisions among those that choose (and do not sell) the female candidate in the first gender decision they made. We then replicate this exercise for those who choose (and do not sell) the female candidate in their second gender decision. Importantly, recall that we randomized the order of these gender decisions. These exercises yield a lower bound on implicit discrimination of 31.0% and 21.8%, with associated p-values of 0.013 and 0.059, respectively (recall it was 29.2% in our main specification with p-value of 0.052). While these results provide strong support for our main finding, the slightly lower male bias among those strictly choosing the female candidate in their second gender decision suggests that balancing may play some role.

Implicit discrimination against women only? As a final robustness check, we conduct a form of placebo exercise in which we replicate our main analysis except that we consider implicit bias according to the two certificates (word or knowledge) instead of gender. This exercise also helps to address the potential concern that our main result might be noise-driven. To do this, we use these two certificate decisions in a similar way to the gender decisions, clas-

sifying employers into explicit certificate bias types. We then examine the implicit certificate bias of each of these types in the complex decisions. If our main findings were due to noise (as opposed to gender information), we would expect also to identify implicit discrimination in this exercise.

Table 4: Empirically identified bounds on implicit discrimination.

	Frequency implicit discriminators (%)		
Attribute of CV	Lower bound	Upper bound	p-value
Female gender	29.2	62.5	0.052
Male gender	-28.8	31.2	0.99
Word certificate	-56.6	15.5	1
Knowledge certificate	-54.8	17.9	1

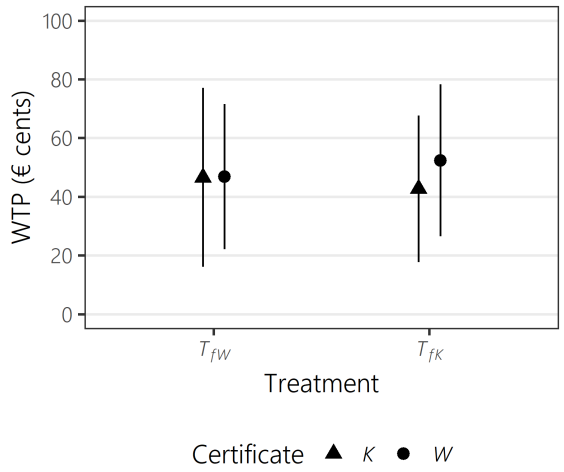
Notes: The p-value is based on a test of the null hypothesis that the lower bound is zero or negative against the alternative that it is greater than zero. The frequency refers to the proportion of employers who decide against the respective attribute in the complex decision (initial choice) among those who explicitly favor it in the gender or certificate decisions, respectively.

Table 4 shows the resulting lower and upper bounds for implicit discrimination by gender (against women and men) and by certificate (against the word and knowledge certificate). We find that the only category in which we observe evidence for implicit discrimination is against women, namely our main finding. We do not find evidence for implicit discrimination against men, nor do we find it in our placebo exercises that examine implicit discrimination against the two certificates. This indicates that noise is highly unlikely to be the main factor driving our findings.

E Redefining Merit

Figure 5 reports employers' willingness to pay (WTP) for a candidate with a specific certificate in comparison to a randomly drawn candidate in each treatment (see footnote 16 for details). Visually, this shows that the mean WTP for the two certificates in treatment T_{fW} are close to one another, while in treatment T_{fK}, the word certificate (which is held by the male candidate here) has a higher mean WTP. We investigate whether there is a statistically significant treatment effect on these beliefs by estimating an interaction effect of treatment and certificate in an OLS regression (standard errors clustered by employer). The point estimate of the interaction coefficient is 9.0 € cents, suggesting that employers have a higher WTP for the word certificate relative to the knowledge certificate in the treatment where the male candidate holds the word certificate ($t_{\text{df=224}} = 1.84, p = 0.067$).

Figure 5: Beliefs about qualification informativeness (by treatment).



Notes: (i) The figure reports the elicited willingness to pay (WTP) of employers to hire a candidate that has a knowledge (*K*) or word (*W*) certificate in comparison to a randomly drawn candidate. (ii) Sample means with ± one sample standard deviation are reported.

One common issue with using a WTP measure is that there tend to be some individuals who engage in multiple switching (Yu et al., 2021). We also observe this in our data. Since we also included a second, simpler measurement of participants' beliefs about the value of each certificate, we use those responses to impute the missing data for those participants who switched multiple times in the price lists. (This occurred in 42 out of the 2·240 = 480 responses, of which 23 [19] are in treatment T_{fW} [T_{fK}], and 20 [22] are for the knowledge [word] certificate.)[54] This exercise alleviates the concern that the missing values bias the estimate, as a similar point estimate for the effect size is found, namely that employers' WTP is 9.3€ cents higher for the

[54]To do this, we exploit the strong theoretical and empirical correlation between the responses in the price-list-based and unincentivized elicitation, which asked about the informativeness of each of the qualifications for performance in the job task on a 5-point scale ($r = 0.48$, $p < 0.001$; see Figure 6 in Appendix G). Specifically, we use multiple imputations based on predictive mean matching and then pool point estimates and variances across analyses of 100 imputed data sets (Rubin, 1987; van Buuren, 2018).

Table 5: Treatment effects on willingness to pay for qualifications.

	Full sample		By discrimination types (Multiple switchers excluded)			
	Multiple switchers excluded	Multiple switchers imputed	Explicit: against men	Explicit: against women	Explicit: mixed	Explicit: non
	(1a)	(1b)	(2a)	(2b)	(2c)	(2d)
T_{fK}	−3.2	−4.0	−22.9**	−4.0	−1.8	2.6
	(3.8)	(3.8)	(10.2)	(8.0)	(6.9)	(6.7)
Word	0.8	0.0	0.0	−6.2	−6.2	11.7*
	(3.8)	(3.8)	(12.5)	(8.7)	(5.7)	(6.0)
T_{fK} × Word	9.0*	9.3*	11.7	23.1**	4.7	4.5
	(4.9)	(4.9)	(14.1)	(10.3)	(8.3)	(8.1)
Constant	45.8***	47.0***	55.0***	45.0***	51.0***	38.1***
	(3.0)	(3.0)	(7.4)	(6.7)	(5.4)	(5.1)
Observations	438	480	59	109	135	135
R^2	0.017	0.015	0.109	0.071	0.006	0.077
RMSE	27.00		26.98	27.30	26.41	25.38

Notes: (i) "T_{fK}" refers to Treatment T_{fK} and "Word" refers to the word certificate. (ii) Every employer completed price lists for each of the two certificates, and treatment effects on the relative valuation of the certificates are therefore reflected in the interaction term. (iii) The unit of the outcome variable is € cents. (iv) Standard errors are clustered at the employer level. (v) Model in column (1b) is based on pooling analyses of 100 imputed data sets where willingness to pay was imputed via predictive mean matching based on the response about certificate informativeness on a 5-point scale. * p < 0.1, ** p < 0.05, *** p < 0.01.

word certificate when it is held by the male candidate ($t_{\text{df=219.18}} = 1.90$, $p = 0.058$; Table 5 reports full regression results).

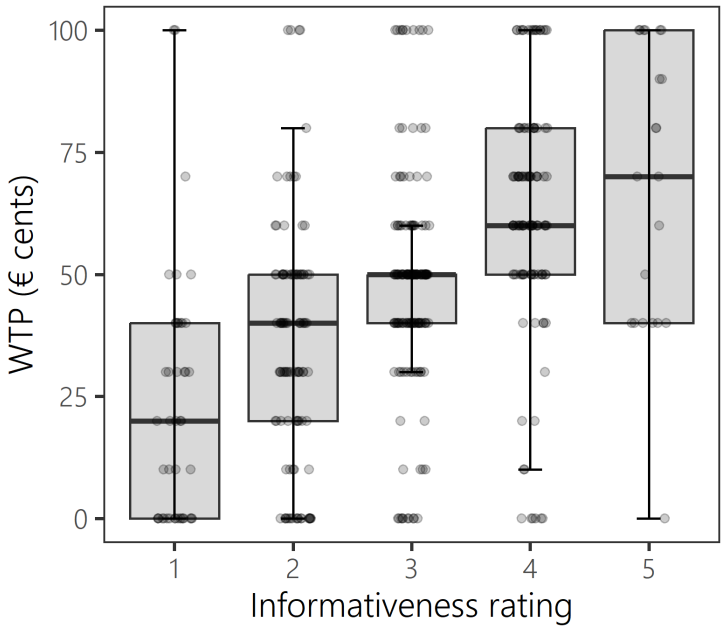
Furthermore, we find that it is exactly the two types of employers that explicitly discriminate by gender whose valuations of the two certificates are subject to treatment effects, thus both ex-post redefining merit depending on which certificate the male candidate had in the preceding complex decision – see Appendix Table 5, columns 2a–d. In particular, even those employers that explicitly discriminate against women appear to justify their predominantly male hiring (in the gender decisions, but also in the complex decision, see main text Table 2), as qualification-based, and we confirm these findings also when revisiting these types' certificate decisions.[55]

Overall, these findings lend additional support to the basic idea of implicit bias arising from gender stereotypes. Facing difficulty in ranking qualifications, employers' belief formation about their value appears to be unconsciously biased by gender stereotypes. Interestingly, this seems to occur both for those employers that would explicitly/consciously discriminate in line with the stereotype and those that would explicitly/consciously act against it. Hence, also our employers explicitly discriminating against women, as via a male bias when candidates are

[55]Explicit discriminators against men, on average, attach the same value to both certificates in treatment T_{fW} (55€ cents), but a lower value to each of them in treatment T_{fK}; and while their valuations of the two certificates in treatment T_{fK} do exhibit quite a gap (32€ cents for K vs. 44€ cents for W on average), the corresponding interaction coefficient is not statistically significant. In contrast, for employers that discriminate explicitly against women, this interaction term is twice the size and significant. Pooling the two decisions Cer_f and Cer_m, both explicitly gender discriminating types hire the knowledge and word candidates at the same (equal) ultimate rate in Treatment T_{fW} (50.0% and 51.7% word hiring), whereas both show a similar and sizable word bias in Treatment T_{fK} (63.3% and 64.8% word hiring), where the male had the word certificate in the initial decision.

equally qualified, may not be fully aware of their discrimination against women at the time of their complex decisions.

Figure 6: Association between beliefs from different elicitation methods.



Notes: Box plot comparing incentivized (WTP via price lists) and non-incentivized (5-point scale from 1="not informative" to 5="very informative") elicitation of certificate informativeness among the same employers. Responses for both certificates are pooled.

F Statistical Accuracy of Discrimination

The empirical economics literature on discrimination has largely focused on the question of whether observed discrimination is mainly taste-based or statistical, as these have different welfare and policy implications. In a recent survey of this literature, Bohren et al. (2022) point out that only very few studies have considered the possibility of *statistically inaccurate* beliefs, and they show that allowing for inaccurate beliefs results in an identification problem. At a general level, this result highlights once again how challenging it is to identify discrimination with naturally occurring data, which prompted the seminal use of field-experimental methods in economics by Bertrand and Mullainathan (2004).

We exploit the additional control afforded by a lab experiment to rule out taste-based discrimination by design and—as demonstrated—be able to identify discrimination without imposing any assumptions on subjective beliefs. It is indeed a distinctive feature of our design that it offers a clear (gender-blind) non-discrimination benchmark by presenting individuals with carefully designed related decisions. Thus, our analysis zooms in on belief-based discrimination, whilst allowing for any subjective and heterogeneous beliefs. This approach has the substantial advantage of not relying on strong assumptions about how employers form beliefs, which is a serious concern especially in laboratory settings. Nonetheless, given the importance of ("accurate") statistical discrimination in the literature, we can also consider how observed discrimination and gender bias relate to actual performance differences in our experiment.[56]

From a theoretical point of view, unless gender information is subjectively perceived to be independent of performance, non-discrimination is often a mistake from the perspective of classical rational decision-making. It means ignoring payoff-relevant information contained in the gender signal. Our employers had monetary incentives to discriminate whenever they believed there were even small performance differences between candidates. Nonetheless, around 13% of all employers are perfectly consistent with non-discrimination across all of their nine decisions (by treatment, 14% and 11%).[57] Examining actual performance differences also allows us to gauge to what extent such non-discrimination is really "irrational" here.

F.1 Performance differences by gender and stereotypes

We first look at the distribution of candidates' actual performance in the job task by gender. The smoothed distribution of scores is shown in Figure 8, and Table 6 adds standard descrip-

[56] In contrast to taste-based discrimination, (accurate) statistical discrimination is often considered efficient, justifiable, or even "fair;" e.g., see the survey by Bertrand and Duflo (2017) for some discussion. Since we rule out taste-based discrimination here, it is worthwhile pointing out that, though individually rational, even accurate statistical discrimination may well be socially inefficient; see Coate and Loury (1993) for seminal work, Fang and Moro (2011) for a survey, and Lepage (2021a,b) for very recent contributions. Moreover, in any case, it violates meritocratic principles.

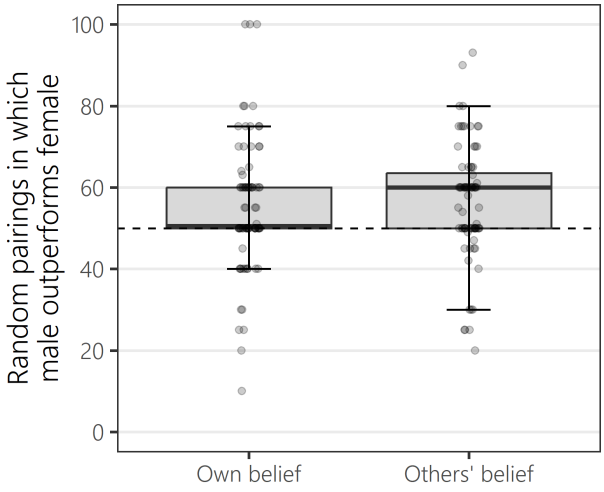
[57] In addition to explicitly not discriminating in both gender decisions, being consistent with non-discrimination across the nine decisions requires choosing consistently between qualifications K, W and – in all other decisions. Only in that way they clearly reveal that gender never matters to them. Under random choice, the probability of satisfying this standard is less than 0.5%.

tive statistics. Though the distributions' shapes differ, we find no evidence of a performance difference between men and women in terms of their mean scores ($t_{\text{df=78}} = 0.23$, $p = 0.82$); moreover, all three quartiles are identical. In fact, if we use the very statistic on which our employers' incentives are based and compare the performance of a randomly drawn man and woman in our pool of job candidates, the probability that the woman scored higher equals 0.53 (see the final row of Table 7, which is discussed in more detail below). Loosely speaking, a discriminating employer that deems qualification in the form of our certificates completely uninformative should therefore always hire a woman over a man to maximize the expected payoff.

This indicates that employers, and specifically those that explicitly discriminate, tend to hold an inaccurate stereotype that men are better at the job task. We are further able to support this conclusion, by drawing on the incentivized beliefs we elicited from the candidates themselves, who are from the same subject pool as employers, as part of the Job Candidate Assessment. On average, our job candidates reported believing that in 55.4 (SD: 16.2) out of 100 comparisons between a randomly drawn man and a randomly drawn woman, the former would perform better in the job task. This is significantly larger than 50 ($t_{\text{df=79}} = 2.99$, $p = 0.0037$; Figure 7). In addition, candidates were asked about their second-order beliefs and also here expressed a clear expectation that others would expect men to be better in the job task, winning on average 56.3 out of 100 random pairings ($t_{\text{df=79}} = 3.98$, $p < 0.001$). Such second-order beliefs can be a sign of persistent stereotypes (Dustan, Koutout, and Leo, 2022). Both response distributions are visualized as boxplots in Figure 7.

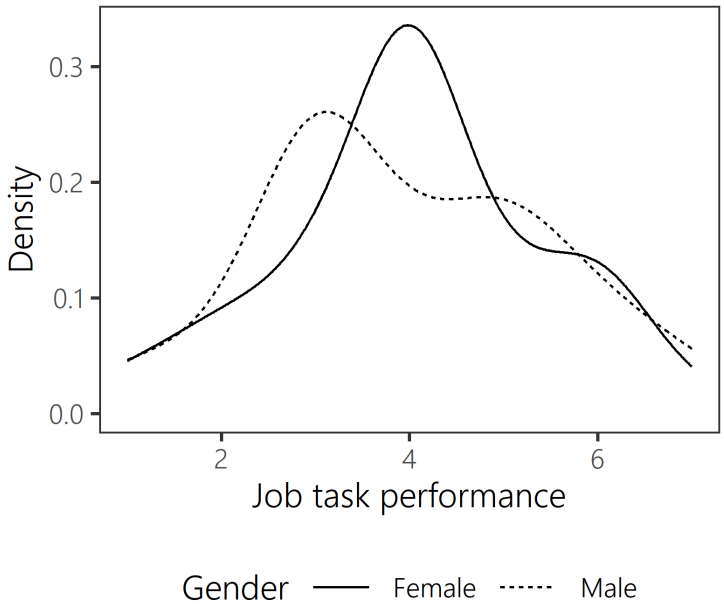
In terms of gender differences in these beliefs, in the first-order beliefs, male candidates held higher beliefs about the performance of men in the job task (mean: 59.4, SD: 14.6) in comparison to the beliefs of female candidates (mean: 52.1, SD: 16.9; $t_{\text{df=78}} = 2.04$, $p = 0.045$ for the comparison). For the second-order beliefs, this gap was smaller and not statistically significant ($t_{\text{df=78}} = 1.19$, $p = 0.24$), with male candidates expecting that others reported a belief of 58.4 on average (SD: 12.5) and female candidates reporting that others reported a belief of 54.6 (SD: 15.3) on average. Since the second-order beliefs are also about individuals of the other gender (candidates were inevitably aware that in their experimental sessions there were participants of both genders), the small reduction in the gap could suggest that candidates are aware to some degree of how their beliefs differ from those of the other gender.

Figure 7: Gender stereotypes about the job task.



Notes: Distribution of responses among job candidates who were asked what they believe is (i) the number of times out of 100 that a random male from their session outperforms a random female candidate in the job task (left boxplot) and (ii) the average belief among the other candidates in the session (right boxplot). Dashed line at 50 indicates the benchmark of perceived gender neutrality of the task.

Figure 8: Job task performance by gender.



Notes: (i) Kernel density estimate for the number of solved matrix exercises by gender of the job candidates.

However, this is only suggestive, since our employers are paid according to whether the candidate they hire performed better than the one not hired, and the hiring decisions they face concern candidate profiles that contain additional, potentially relevant information. There-

Table 6: Descriptive statistics on performance in the job task (number of solved matrix exercises) by gender .

Statistic	Female	Male
N	44	36
Mean	4.05	3.97
SD	1.38	1.52
Min	1	1
25% Pctl	3	3
Median	4	4
75% Pctl	5	5
Max	7	7

fore, in a specific hiring decision, what is relevant for payoff maximization is the performance comparisons *conditional* also on this additional information. The next section examines these conditional comparisons.

F.2 Statistical accuracy of discrimination and non-discrimination

Table 7 considers each of the hiring decisions and reports the conditional probability that a randomly drawn candidate with the characteristics of Candidate A performed better than a randomly drawn candidate with the characteristics of Candidate B, with ties broken randomly (i.e., the tie probability mass is distributed equally on the two candidates, exactly as the random tie-breaking in determining employers' payoffs). Rows 1–10 tell us who should be chosen in each of the ten decisions to maximize expected earnings according to the true conditional performance distributions (it is optimal to sell if the probabilities are less than 0.5167, see also footnote 14). Additionally, rows 11–13 of the table carry out the analogous calculation for "gender-blind" comparisons that would be relevant to non-discriminators. The final row 14 shows the comparison of a randomly selected female candidate and a randomly selected male candidate.

It is clear from the table that subjective beliefs that consider qualification information irrelevant would be inaccurate. In fact, from a purely statistical point of view, both gender and certificates are informative about job performance; in other words, non-discriminators forgo earnings.

Accurate statistical discrimination would indeed result in a gender bias against women in the complex hiring decisions, and even in the simple ones (due to decision Sim_{fW}, where, despite being more qualified, a randomly drawn woman with a word certificate in our sample is no likely to perform better on the job task than a randomly drawn man). However, the gender bias we actually observe is mainly due to (explicit discriminators') preferential hiring of men in Treatment T_{fK}'s corresponding complex decision, where there truly is no performance difference. Moreover, in contrast to the similar degrees of gender bias observed against women in our data in both gender decisions, accurate statistical discrimination would imply no gender

bias upon aggregating over the two. Hence, our explicit discriminators—i.e., those employers even openly relying on gender information—tend to be statistically inaccurate overall. Indeed, not a single employer in our experiment succeeds in consistently maximizing expected earnings across all their nine decisions.

Table 7: Observed probability for being the better candidate in a comparison of given CVs (with ties broken randomly).

	A	B	Pr(A better)	Pr(B better)	Pr(tie)
Com_{fW}	(f,W)	(m,K)	0.426	0.574	0.121
Com_{fK}	(f,K)	(m,W)	0.499	0.501	0.169
Gen_K	(f,K)	(m,K)	0.554	0.446	0.171
Gen_W	(f,W)	(m,W)	0.374	0.626	0.107
Cer_f	(f,W)	(f,K)	0.356	0.644	0.313
Cer_m	(m,W)	(m,K)	0.540	0.460	0.188
Sim_{fK}	(f,K)	$(m,-)$	0.633	0.367	0.164
Sim_{fW}	(f,W)	$(m,-)$	0.500	0.500	0.139
Sim_{mK}	$(f,-)$	(m,K)	0.456	0.544	0.148
Sim_{mW}	$(f,-)$	(m,W)	0.406	0.594	0.139
n.a.	$(-,K)$	$(-,W)$	0.542	0.458	0.190
n.a.	$(-,K)$	$(-,-)$	0.578	0.422	0.188
n.a.	$(-,W)$	$(-,-)$	0.536	0.464	0.199
n.a.	$(f,-)$	$(m,-)$	0.528	0.472	0.161

We can conduct an analogous statistical exercise for those employers that are consistent with non-discrimination, by examining whether they maximize expected earnings but subject to a gender-blindness constraint. As Table 7 shows in rows 11–13, either certificate indicates better performance than that of a purely random candidate (corresponding to $(-,-)$ in the table) and, when ignoring gender information, statistically, the knowledge certificate should be favored over the word certificate. However, when we consider the set of employers that are consistent "non-discriminators"[58] across all of the decisions where both candidates have a certificate—i.e., the complex, gender and certificate decisions—we find that around 64% hire the candidate with the word certificate in the two certificate decisions (64.3% and 63.9% in Treatments T_{fW} and T_{fK}, respectively, and 62.9% and 64.1% in terms of ultimate as opposed to initial hiring). In this sense, the majority of gender non-discriminators with consistent hiring patterns could still be classified as statistically inaccurate.

When it comes to reducing discrimination via information campaigns, as advanced by Bohren et al. (2022), inaccurate non-discriminators are a key target group, however: Any gender bias they cause is due to inaccurate beliefs about qualifications only, and it will disappear if

[58]These are employers who engage in one of two patterns of consistent non-discriminatory (gender-blind) behavior. Either they (i) sell their choices in all five decisions, or (ii) they sell their choices in the two gender decisions and always follow the candidate with the same certificate in the other three decisions where the candidates hold different certificates, namely the complex and certificate decisions.

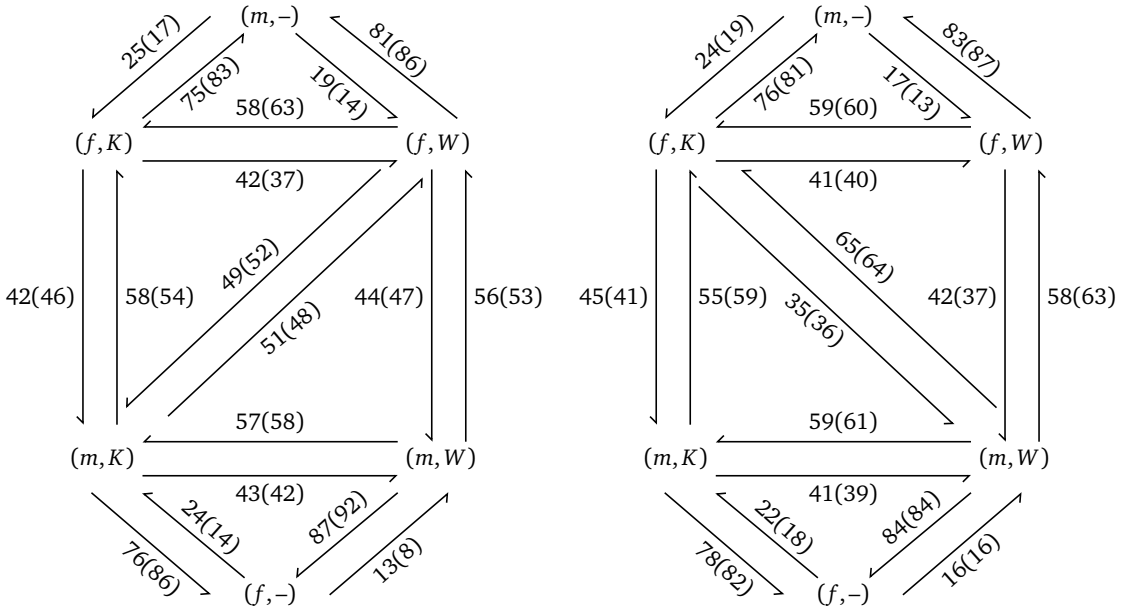
their beliefs are corrected. For example, if qualifications that men are more likely to have tend to be overvalued, even those committed to non-discrimination introduce an unwarranted gender bias in favor of men due purely to their mistaken beliefs about the value of qualifications more likely to be held by men. For rational discriminators, information campaigns will work to reduce discrimination only in settings where there truly are no group differences; otherwise such campaigns may even backfire.

G Additional Tables and Figures

Table 8: Demographic information of participants in the HIRING EXPERIMENT.

Variable	Treatment T_{fW} $(N = 118)$	Treatment T_{fK} $(N = 122)$
Age (mean, SD)	24.8 (5.4)	24.2 (4.4)
Gender: female (N, %)	58 (49.2)	61 (50.0)
Study subject: STEM (N, %)	59 (50.0)	65 (53.3)
Study subject: Econ./Business (N, %)	39 (33.1)	40 (32.8)

Figure 9: Ultimate (initial) choice propensities in both treatments



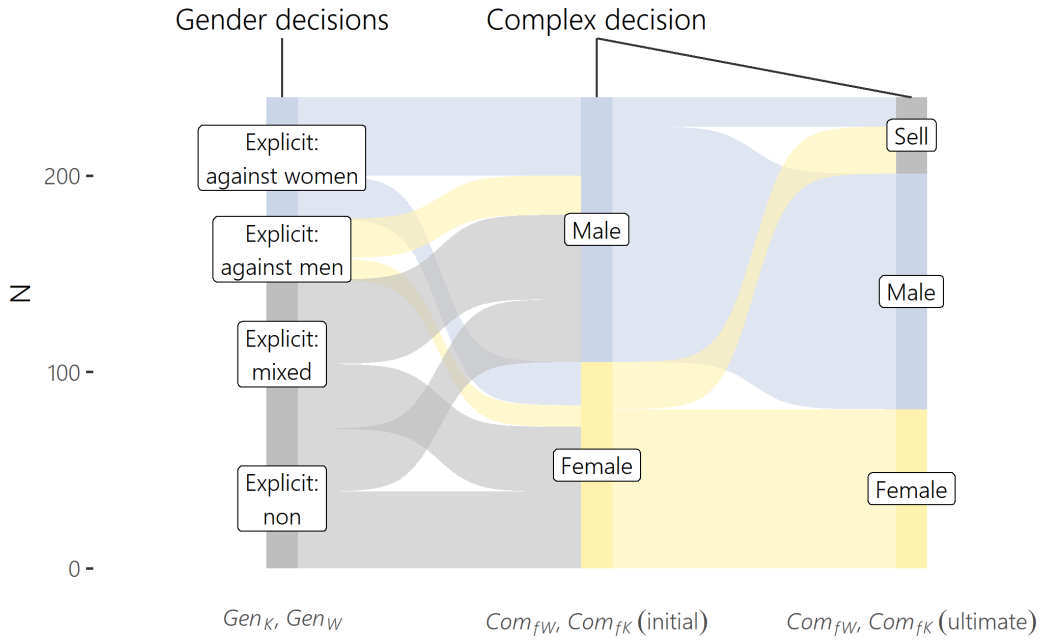
Notes: (i) The left-hand panel summarizes the aggregate choices from Treatment T_{fW} $(N = 118)$, while the right-hand panel describes the same information for Treatment T_{fK} $(N = 122)$. (ii) Each edge of the figure contains a pair of parallel arrows and corresponds to a binary decision between two candidates (g, C) and (g', C'), which are described at the nodes of the figure (see Table 1). Each arrow is associated with two propensities, $X(Y)$, associated with the initial and ultimate choice propensities. This means that for an arrow from (g, C) to (g', C') that X% (resp., Y%) of employers ultimately (resp. initially) hire (g, C) over (g', C'). The parallel arrow contains the corresponding propensities, such that summing the propensities for pairs of arrows always equals 100.

Table 9: Aggregate results from all hiring decisions by treatment.

			Treatment T_{fW}			Treatment T_{fK}		
Decision	A	B	Pr. hire A init.	Pr. sell	Pr. keep A	Pr. hire A init.	Pr. sell	Pr. keep A
Com_{fW}	(f,W)	(m,K)	0.517 (61/118)	0.178 (21/118)	0.398 (47/118)	–	–	–
Com_{fK}	(f,K)	(m,W)	–	–	–	0.361 (44/122)	0.148 (18/122)	0.279 (34/122)
Gen_K	(f,K)	(m,K)	0.458 (54/118)	0.424 (50/118)	0.212 (25/118)	0.410 (50/122)	0.393 (48/122)	0.254 (31/122)
Gen_W	(f,W)	(m,W)	0.466 (55/118)	0.483 (57/118)	0.195 (23/118)	0.369 (45/122)	0.410 (50/122)	0.213 (26/122)
Cer_f	(f,W)	(f,K)	0.627 (74/118)	0.178 (21/118)	0.492 (58/118)	0.598 (73/122)	0.156 (19/122)	0.516 (63/122)
Cer_m	(m,W)	(m,K)	0.576 (68/118)	0.161 (19/118)	0.492 (58/118)	0.615 (75/122)	0.189 (23/122)	0.500 (61/122)
Sim_{fK}	(f,K)	$(m,-)$	0.831 (98/118)	0.322 (38/118)	0.585 (69/118)	0.811 (99/122)	0.303 (37/122)	0.607 (74/122)
Sim_{fW}	(f,W)	$(m,-)$	0.856 (101/118)	0.271 (32/118)	0.678 (80/118)	0.869 (106/122)	0.180 (22/122)	0.738 (90/122)
Sim_{mK}	$(f,-)$	(m,K)	0.144 (17/118)	0.339 (40/118)	0.068 (8/118)	0.180 (22/122)	0.230 (28/122)	0.107 (13/122)
Sim_{mW}	$(f,-)$	(m,W)	0.085 (10/118)	0.229 (27/118)	0.017 (2/118)	0.164 (20/122)	0.197 (24/122)	0.066 (8/122)

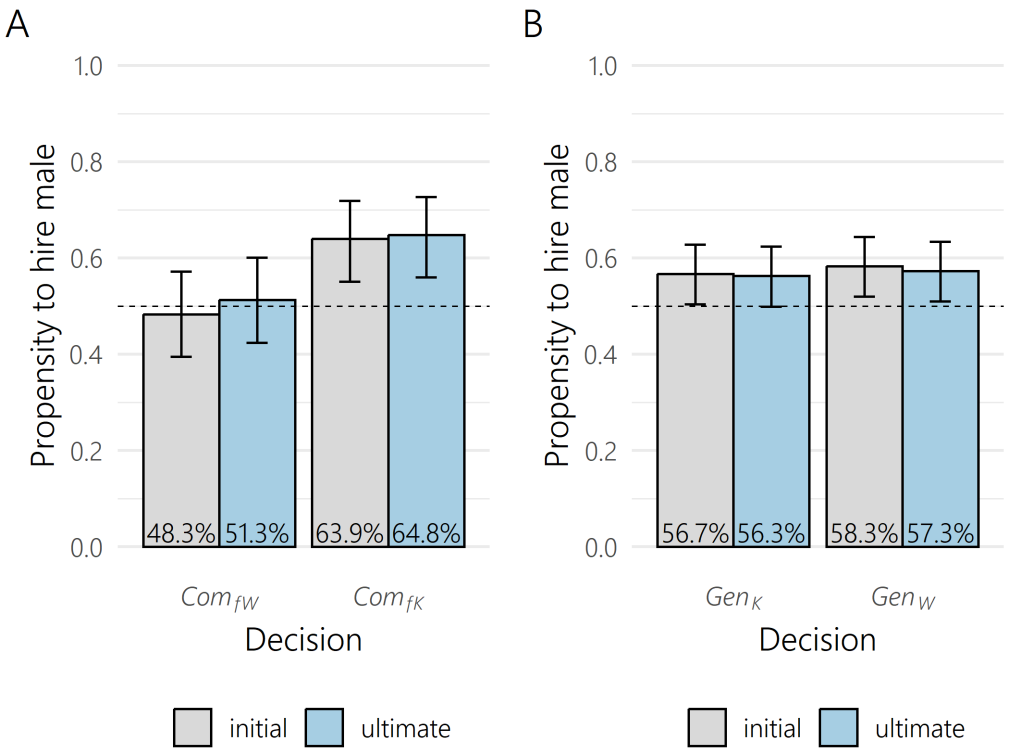
Notes: "Pr. hire A init." refers to hiring candidate A initially, "Pr. sell" refers to selling the initial choice (regardless of whether it was A or B), and "Pr. keep A" refers to hiring candidate A initially and not selling this initial choice. Hence, the difference between "Pr. hire A init." and "Pr. keep A" equals the fraction of employers that hire candidate A initially and then sell this initial choice; combining this with "Pr. sell" yields the fraction of employers that hire candidate B initially and then sell this other initial choice. For instance, in complex decision Com_{fW}, $(51.7-39.8)\% = 11.9\%$ initially hire the woman and then sell the choice, and $(17.8-11.9)\% = 5.9\%$ initially hire the man and then sell this choice; this means that conditional on hiring the woman the selling rate equals 23.0%, whereas it equals 12.3% conditional on hiring the man.

Figure 10: Identifying explicit and implicit discrimination from within-subject data.



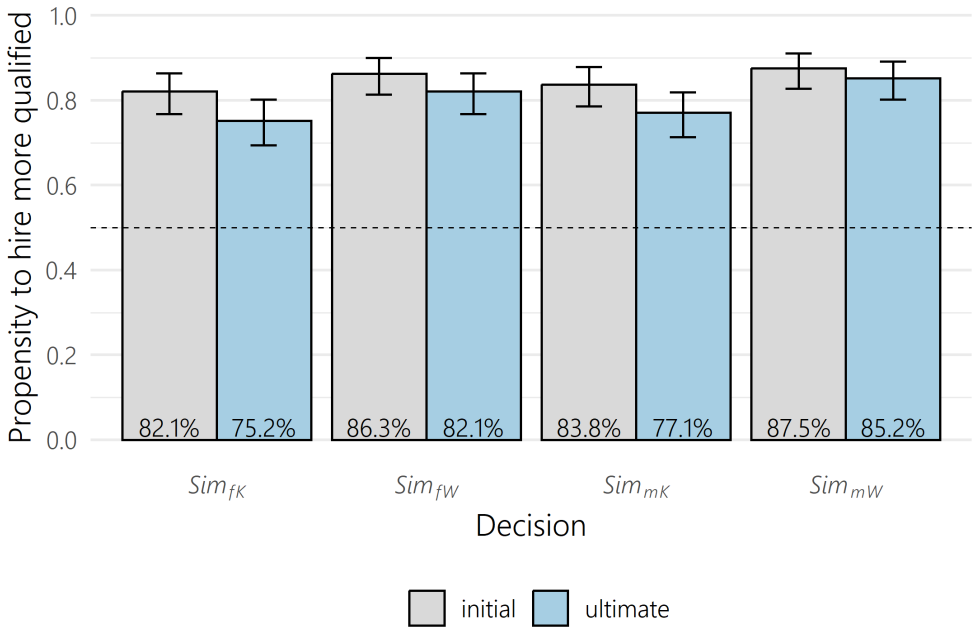
Notes: (i) The figure shows the relationship between the classification of employers into explicit discrimination types, as defined in the gender decisions, and their decision-making in the complex decision. (ii) The colors of the flows and columns allow us to track the decisions of the employers in favor of the male (blue) or female (yellow) candidate across different choice settings. (iii) The two treatment groups are pooled together in this figure. (iv) The y-axis reports the number of employers for scale.

Figure 11: Hiring choices in complex (A) and gender (B) decisions.



Notes: (i) Panel A reports the propensity to hire the male candidate in the two complex decisions. (ii) Panel B reports the propensity to hire the male candidate in the two gender decisions. (iii) Error bars show 95% confidence intervals. (iv) The dashed horizontal line indicates the non-discrimination benchmark, namely hiring men and women at the same rate.

Figure 12: Hiring choices in simple decisions.



Notes: (i) The figure shows the propensity to hire the more qualified candidate in the four simple decisions, where the gender and certificate of the more qualified candidate varies. (ii) As described in Table 1, in decisions Sim_{fK} and Sim_{fW} the more qualified candidate is female, with a knowledge certificate in the former and a word certificate in the latter; analogously, in decisions Sim_{mK} and Sim_{mW} the more qualified candidate is male, with a knowledge certificate in the former and a word certificate in the latter. (iii) In all four decisions, the comparison is between a male and a female candidate. (iv) Error bars show 95% confidence intervals. (v) The dashed horizontal line indicates an equal aggregate propensity to hire both candidates in a particular choice.

H English Translation of the German Instructions

Note: This is the plain text translation of the original German instructions. It includes the original colors as shown on the participants' screens. These instructions omit the candidates' CVs, interactive buttons that participants could click on and the fields where they could enter written text etc. [an example of a decision page in which employers made a decision between two candidate's CVs is portrayed in the screenshot which we have included in the main text of the paper]. A vertical line indicates a new page or part of a page that popped up after a user completed a task or pressed a button.

This is a decision making study. Thank you for your participation. As part of this study, you can earn money that will be paid to you in cash at the end of the experiment. The experiment will last approximately **60 minutes**.

You will receive € 5 for showing up on time. In addition, you will be paid your **earnings** from the experiment. During this experiment you will make several decisions and your additional earnings will depend on these decisions. Therefore, before each decision you will be informed about how it will affect your earnings. None of your decisions can lead to losses.

The **anonymity** of all your decisions is guaranteed. It will not be possible for anyone to associate your identity with the choices you make here.

Please observe the following **ground-rules**:

You are **not** allowed to use electronic devices or to communicate with other participants during the experiment. Please use only the programs and functions intended for the experiment. Please do not talk to the other participants. If you have a question, please raise your hand. We will then come to you and answer your question silently. Please do not ask your questions out loud under any circumstances. If the question is relevant to all participants, we will repeat it out loud and answer it. If you violate these rules, we will have to exclude you from the experiment and the payout.

On the individual pages of this experiment, a **time limit** is displayed at the top of the screen. This is intended as a guideline, so you will **not** be automatically redirected to the next page after it has expired. Nevertheless, please try to keep to the time limit.

On the next page you will receive a short introduction to today's experiment. This contains the **background information necessary** for your later decisions. Please note that the decisions themselves are not particularly time-consuming, but it is all the more important that you **read the following background information carefully** beforehand.

Now click on the button when you are ready! (Please note that at some points in the experiment you will have to wait until all participants have finished before continuing. We ask for your patience while waiting in this case).

In today's experiment, you will take on the role of an **employer**. During the experiment you will face a series of decisions. In each decision, you will choose which of two people you would like to select for a task (i.e., hire for a job). Your **earnings** will depend on whether you select the person who is **better** at this job task. To inform your decision, you will receive a **short profile** with information about each person.

Who are the candidates?

These are 80 actual participants who took part in a previous experiment. These 80 people worked on a series of tasks, specifically in the areas of **logical reasoning**, **knowledge**, and **words**. Each person was paid for each correct answer, so they all had an incentive to perform as well as possible in each task. The individual performance of these past participants in the tasks will be relevant for your decisions and earnings. For details on the tasks that these previous participants completed, please refer to the **printout on your table**. You now have ample time to **familiarize yourself** with these tasks in order to make better decisions later.

What is the objective when choosing who to hire?

Your goal as an employer is to hire one person from each pair of candidates who **performed better** in the **Logical Reasoning** task (when the two candidates achieved the same score in the **Logical Reasoning** task, ties are broken randomly). In other words, logical reasoning is the job task, and you as an employer are interested in selecting the person who did the job best, because that is what your own earnings is based on, as is common for employers.

What information about the candidates will you receive?

When you make your decision, information about the two candidates that you will select between will be available in the form of a short profile for each candidate. More specifically, the profiles include information about whether the person has a **certificate** in **knowledge** or in **words**.

What is a certificate?

A certificate in **knowledge** indicates that the person is among the top 30% of all participants in the **knowledge** task. Since there were a total of 80 participants, you will know that a person with a knowledge certificate has achieved **one of the best 24** performances in the **knowledge** task (ties are decided at random).

A certificate in **words** indicates that the person is among the top 30% of all participants in the **words** task. Since there were a total of 80 participants, you will know that a person with

a words certificate has achieved **one of the best 24** performances in the **words** task (ties are decided at random).

Conversely, the absence of a certificate accordingly means that you do not know for this person whether he or she belongs to the top 30% in the corresponding task or not. In this case, this remains **uncertain**.

Certificates are displayed in the profiles as follows. A check mark in a green field means that the person has a corresponding certificate, i.e. the person belongs to the top 30% in this task **with certainty**:

A question mark in a red field means that there is no corresponding certificate for the person, so it remains **uncertain** for you whether the person belongs to the top 30% in this task:

How are your earnings calculated?

You will receive €6 for your decision if you choose the candidate who achieved a higher performance in the **Logical Reasoning** task. Otherwise you will receive €0.

Keep in mind that all the candidates that you will have to choose between **actually exist** and participated in our previous experiment. This also means that there may be **several real individuals** matching the description that you see in a given profile. In this case, one of these matching candidates will be **randomly chosen by the computer**.

You will also have **additional opportunities** to earn money in the experiment and will be informed about the details when you get there.

A brief summary:

- Your task is to select the candidate who achieved the higher performance in the **Logical Reasoning** task.
- You will receive information about each candidate available for selection. In particular, you will receive information about whether they have a certificate in **knowledge** or **words**.

Notes:

- Your decisions today affect only your own payoff. They will have absolutely no influence on the individuals that we present to you as job candidates. These previous participants have already been paid for their performance in the earlier experiment based on the correct answers they achieved.
- These previous participants also did not know that they would be considered as job candidates in today's experiment and their identity is kept completely anonymous. They were also not informed whether they had achieved a certificate or not.

In a few seconds the button to start the experiment will appear (see above).

Which person do you select for their Logical Reasoning? ***[Answer Options: Person A or Person B]***

You will receive €6 if you choose the person who has a higher score in the **Logical Reasoning** task. Otherwise you will receive €0.

If the employer chooses Person A.

You have chosen Person A.

Would you prefer to instead leave your decision between the two candidates to chance?

Would you like to let the decision that you just made be replaced by a random selection by the computer between the two candidates and **receive €0.10**?

If you choose "No", then you will**keep** your choice of candidate, and, as already explained, you will be rewarded with €**6** if this person performed better, and with €**0** if not.

If you choose "Yes", then you will receive €**0.10 for sure**. The candidate that you selected will be replaced by a **random selection** by the computer between the two candidates. If the computer selects the better candidate, then you will be rewarded with €**6** in addition to the €**0.10** you already received, and with €**0** if not.

Would you prefer to instead leave your decision between the two candidates to chance? ***[Answer Options: Yes or No]***

If the employer chooses "No".

You have chosen No.

Now, what would you like to decide in the following situation?

Your previous choice has been finalized. Now you have the opportunity to earn some additional money. A **new person** has been **randomly** drawn by the computer from all 80 participants. You are again interested in their performance in the job task. You will not receive any more information about this person.

You now have the option to replace the randomly drawn person with another person who has a **specific certificate**. This replacement person is then again drawn randomly by the computer, but from the 24 best participants of the task corresponding to the certificate.

You will receive € 3 if the person finally chosen is among the **best 50%** (= **among the best 40 persons**) for the job, i.e. in the **Logical Inference** task (ties are decided at random).

We would now like to know how much the two possible certificates, in **knowledge** and **words** respectively, are worth to you. For each of the two certificates you will get a price list. That is, we propose prices in ascending order. For each price you need to decide whether you would like to replace the person drawn at random from all 80 participants with a person with that particular certificate. You have a budget of € 1 for each decision and must decide for each price whether you would pay this price ("Yes") or not ("No").

When you have made all the decisions for both price lists, the computer randomly selects one of these decisions, which then contributes to your payoff. If you answered "Yes" in this chosen decision, the person initially selected by the computer will be replaced with one with a certificate and you will pay the specified price. You will then receive the balance of the **€ 1**, and an additional **€ 3** if the selected person with certificate is among the best 50% for the **Logical Inference** job. On the other hand, if you answered "No" in this decision, no replacement will be made. You will receive the full **€ 1**, and an additional **€ 3** if the person originally selected from all 80 participants is among the top 50% for the **Logical Inference** job.

To achieve the highest possible payment, you should make each decision as if it were relevant to your payoff. To do this, go through each of the two lists **from top (lowest price) to bottom (highest price)**, answering "Yes" until you reach the first price that is too high for you. Then answer "No" consistently from that line, all the way to the highest price at the bottom of the list. Thus, the more certain you are that a person **with the appropriate certificate** is more likely to be in the top 50% in the **Logical Reasoning** task than a random person, the more prices you should answer "Yes" for at the beginning of the list before switching to "No".

What is your evaluation of the value of the certificates?

How **valuable** do you consider the certificates to be in predicting the performance of **a given individual** in **Logical Reasoning**? Please indicate your rating of the meaningfulness of the certificates on a scale of 1 to 5, with

1 = **not** meaningful,
2 = **not very** meaningful,
3 = **moderately** meaningful,
4 = **fairly** meaningful,
5 = **very** meaningful.

You will not earn any extra money in this task.
