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

Working Paper

Explicit and implicit belief-based gender discrimination: A hiring experiment

Discussion Paper, No. 325

Provided in Cooperation with:
University of Munich (LMU) and Humboldt University Berlin, Collaborative Research Center Transregio 190: Rationality and Competition

Suggested Citation: Barron, Kai; Ditlmann, Ruth; Gehrig, Stefan; Schweighofer-Kodritsch, Sebastian (2022) : Explicit and implicit belief-based gender discrimination: A hiring experiment, Discussion Paper, No. 325, Ludwig-Maximilians-Universität München und Humboldt-Universität zu Berlin, Collaborative Research Center Transregio 190 - Rationality and Competition, München und Berlin

This Version is available at:
https://hdl.handle.net/10419/256792

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 (HU Berlin)

Discussion Paper No. 325

April 25, 2022

Collaborative Research Center Transregio 190 | www.rationality-and-competition.de
Ludwig-Maximilians-Universität München | Humboldt-Universität zu Berlin
Spokesperson: Prof. Georg Weizsäcker, Ph.D., Humboldt University Berlin, 10117 Berlin, Germany
info@rationality-and-competition.de

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

April 24, 2022

Abstract

Understanding discrimination is key for designing policy interventions that promote equality in society. Economists have studied the topic intensively, typically taxonomizing discrimination as either taste-based or (accurate) statistical discrimination. To reveal the limitations of this taxonomy and enrich it psychologically, we design a hiring experiment that rules out both of these sources of discrimination with respect to gender. Yet, we still detect substantial discrimination against women. We provide evidence of two forms of discrimination, *explicit* and *implicit* belief-based discrimination. Both rely on statistically inaccurate beliefs but differ in how clearly they reveal that the choice was based on gender. Our analysis highlights the central role played by contextual features of the choice setting in determining whether and how discrimination will manifest. We conclude by discussing how policy makers may design effective regulation to address the specific forms of discrimination identified in our experiment.

JEL Codes: D90, J71, D83

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

*The authors would like to thank Vojtech Bartos, Katherine Coffman, 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, and at Stockholm University's SOFI Labour Economics Seminar for helpful comments. We thank the WZB for generously funding this project through means of its interdisciplinary "seed money" programme. Barron and Schweighofer-Kodritsch gratefully acknowledge financial support by the *Deutsche Forschungsgemeinschaft* through CRC TRR 190 (project number 280092119).

1 Introduction

Discrimination in the labor market is a critically important policy issue around the world.[1] When one individual receives preferential treatment over another on the basis of gender or ethnicity, this violates basic meritocratic principles. It is also inefficient if it results in a less productive workforce due, for instance, to (i) lower returns to educational investments for some groups, or (ii) sub-optimal matching between skills and tasks. Moreover, such discrimination predominantly harms socio-economically weaker groups and thereby reinforces inequality. 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 wages and career progression.[2] However, in addition to being able to detect discrimination it is crucial for the design of effective policies to be able to understand the underlying causes of discrimination.

Traditionally, the economics literature has distinguished between discrimination based on taste (Becker, 1957) and discrimination resulting from rational, i.e., statistically accurate beliefs about groups (Phelps, 1972; Arrow, 1973). In the former, an employer is willing to pay a price to avoid transactions or being "associated with some persons instead of others" (Becker, 1957, p. 14). In the latter, employers' discrimination is an optimal response to truly existing productivity differences between groups. However, recent work suggests that this taxonomy may be too narrow in important ways. First, this recent literature emphasizes the prevalence of discrimination emanating from "irrational" (i.e., inaccurate) beliefs due, for example, to widely held stereotypes (see, e.g., Judd and Park, 1993; Hilton and Von Hippel, 1996; Heilman, 2001; Bordalo, Coffman, Gennaioli, and Shleifer, 2016; Bohren et al., 2020; Mengel and Campos-Mercade, 2022). Second, building on insights from social psychology (e.g., Snyder, Kleck, Strenta, and Mentzer, 1979; Banaji and Greenwald, 1995; Hodson, Dovidio, and Gaertner, 2010), it highlights a tension between complying with social norms against discrimination—e.g., to maintain a positive social or self-image—whilst holding discriminatory stereotypes/preferences (see Bertrand, Chugh, and Mullainathan, 2005; Bohnet et al., 2015; Bertrand and Duflo, 2017; Danilov and Saccardo, 2017; Carlana, 2019; Cunningham and de Quidt, 2022). In decisions where judgment is cognitively difficult or very subjective, an individual facing a tension of this nature is prone to discriminate

[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) and Blau and Kahn (2017).

[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, 2020; Coffman, Exley, and Niederle, 2020), *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).

implicitly, contradicting the beliefs/preferences she expresses *explicitly*. Such implicit discrimination falls outside the scope of the (subjective) expected utility framework on which all of the aforementioned explanations are based (see especially Cunningham and de Quidt, 2022). It is a particularly problematic form of discrimination because: (i) by it's 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.

In this paper, we study both explicit and implicit gender discrimination using a stylized labor market hiring experiment. We do this at both the aggregate level, across different types of hiring decisions intended to vary the degree of subjectivity in ranking candidates' qualifications, and the individual level, classifying employers into discrimination types and revealing the heterogeneity in their behavior. The design allows us to completely rule out classical taste-based discrimination and focus on isolating different forms of belief-based discrimination. In this, we join the rapidly growing contemporary literature considering the central role of (possibly inaccurate) beliefs in generating different behavior towards men and women.[3]

To identify both explicit and implicit belief-based discrimination, we designed a tightly controlled experiment that simulates the main features of a hiring scenario, but allows us to study these different manifestations of discriminatory behavior. This requires making careful adjustments to the information environment in which participants make their hiring decisions, such that we observe the same individual making a sequence of hiring choices where the characteristics of the candidates are systematically varied. In the experiment, a first group of 80 participants serve as job candidates, and a second group of 240 participants take on the role of employers. These employers make a series of anonymous hiring decisions between pairs of candidates. For each candidate, the employer receives a "mini-CV" that provides information about 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. Our incentive structure achieves multiple objectives. First, the fixed payment (as opposed to paying the employer

[3]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., 2020; Mengel and Campos-Mercade, 2022; 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. (2020) 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 5 of the 105 papers they identified in top economics journals tested for inaccurate beliefs. Collectively, the recent literature suggests this is an important omission.

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 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. (2020) or Reuben et al. (2014), where the evaluator's/employer's choices directly affect the candidate's fate.[4]

To identify discrimination, we compare decisions made by employers between real candidates with "qualification profiles" a and b, say, where in one scenario a belongs to a female and b to a male candidate, and in another scenario this is reversed. Any significant difference in how often qualification profile a gets chosen between the two scenarios ("gender bias" in hiring rates) can be cleanly attributed 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.

We find the following patterns of belief-based gender discrimination at the aggregate level: First, whenever both candidates are *equally qualified* (i.e., both have identical qualification profiles), there is significant discrimination against women. Since such decisions are only about gender, and 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., 2020), since female candidates were not objectively out-performed by male candidates; in this sense, our experiment rules out (accurate) statistical discrimination. Second, whenever one profile is *more qualified* than the other (i.e., one has a certificate, the other has none), the more qualified candidate is always hired at the same rate, irrespective of gender; i.e., there is no gender bias.

However, third, whenever both profiles are *differently qualified* (i.e., each candidate has a different certificate), there is again a significant gender bias and hence discrimination against women. The magnitude of the gender bias against women here is approximately as large as for the decisions where employers have no information upon which to differentiate the candidates other than their gender. While the comparison between the scenarios where one candidate is more qualified and where candidates are differently qualified suggests

[4]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.

implicit discrimination due to increased subjectivity, the observation that the gender bias in the latter case is actually similar in magnitude to when the hiring decision is explicitly about gender means it could be due to the very same employers that also explicitly discriminate.

Our individual analysis allows us to better understand what is driving these patterns of behavior observed at the aggregate level. We first classify individual employers into explicit discrimination types based on their hiring choices when candidates are equally qualified. 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 (notably, this identification approach is in line with the formal framework for identifying implicit preferences more generally proposed by Cunningham and de Quidt, 2022). While the majority of employers does not display an explicit bias against either gender when candidates are equally qualified, approximately 40% do so. These employers are twice as likely to explicitly discriminate 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 is of very similar magnitude for the two explicit discrimination types. In particular, 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 discrimination is opaque, i.e., implicitly.[5]

The pattern of behavior that we observe is consistent with a stereotype about male superiority in logic tasks leading employers to form gender-biased beliefs (see also Bordalo et al., 2019). In fact, 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), supporting this explanation. The implicit discrimination that we identify therefore appears due to a combination of: (i) an aversion to displaying overtly discriminatory behavior, and (ii) underlying (mistaken) stereotypes. This aversion amongst some participants to overtly discriminate against women may emanate from the fact that gender discrimination against women is stigmatized in certain segments of the population (our subject pool is comprised of young, highly educated Westerners). Despite any beliefs they may hold about who is the better candidate, employers may wish to signal that they do not discriminate. While we cannot unequivocally demonstrate that such self- or social image concerns drive implicit discrimination in our experiment, additional features of our experimental design allow us to study whether employer behavior is consistent with this explanation: Beliefs data we elicit about the relative importance of the two potential qualifications for job performance suggests that employers rationalize gender-biased choices as being qualification-driven by "redefining merit" ex post (see, e.g., Uhlmann and Cohen, 2005). This involves adjusting

[5]By contrast, employers that choose to "non-discriminate" when discrimination would be explicit, also display no gender bias when candidates are differently qualified.

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.

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 of 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.[6] 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."[7] 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 discrimination when candidates are *differently qualified*, but not when one candidate is *more qualified*.[8] 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 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 implicit association test (IAT) in social psychology (Greenwald, McGhee, and

[6]Bohren et al. (2020) 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.

[7]Cunningham and de Quidt (2022) refer to these as *ceteris paribus* rules. We will use a more general terminology and refer to them as "affirmative action rules."

[8]In addition, we also document evidence of discrimination 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).

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).[9] 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 (2022) 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.[10]

The paper proceeds as follows. Section 2 describes the experimental design. Section 3 presents the aggregate level results, Section 4 discusses the individual level analysis, and Section 5 relates our results to objective statistical patterns in the performance data. Section 6 discusses the policy implications and concludes.

2 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 (qualification task 1, which we refer to as the *knowledge task*), a word search puzzle (qualification task 2, 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.

[9]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, Carlana, La Ferrara, and Pinotti (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.

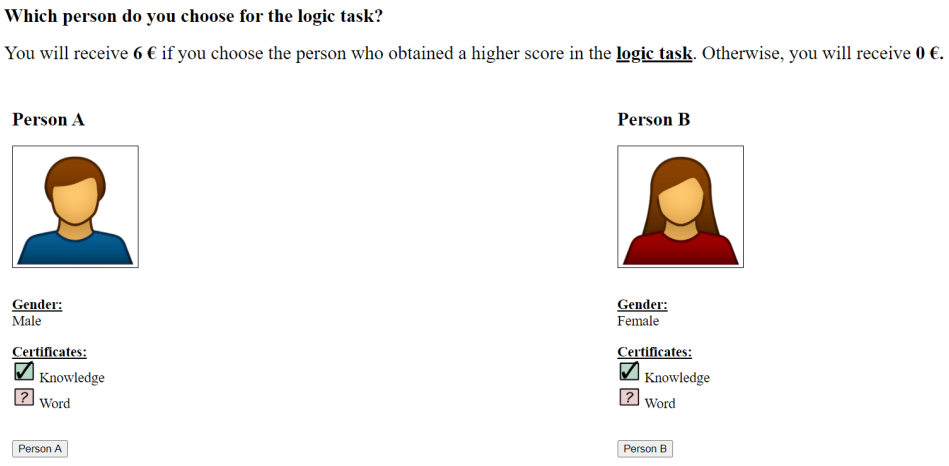
[10]For example, when designing surveys, using the IAT to measure a respondent's implicit biases may be impractical, but adding a few carefully designed (hypothetical) choice questions that vary in how strongly they reveal the decision maker's motives may well be feasible.

2.1 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). 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 decision 2 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 (decision 1), where the choice was between a male and female candidate who were *differently qualified*, two gender decisions between *equally qualified* male and female candidates (decisions 2-3), two qualification decisions between *differently qualified* candidates of the same

gender (decisions 4-5), and four *simple* decisions, where one candidate was *more qualified* (decisions 6-9).[11] Table 1 summarizes the decisions.

A simple hiring decision refers to a choice between a male and a female candidate where one candidate is more qualified, i.e., where one candidate has a certificate whereas the other has none. In the gender and qualification decisions, candidates only differ on one dimension (gender or qualification, respectively). In the complex hiring decision, both candidates are qualified (i.e., have one certificate), but differ on two dimensions (gender and the type of certificate). These complex decisions, therefore, involve a choice between differently qualified male and female candidates.

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 (decision 1). In treatment 1 (T1), the female candidate had a word certificate in the complex decision, and the male candidate had a knowledge certificate. In treatment 2 (T2), 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).

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

	Candidate A		**Candidate B**		
	Gender	**Certificate**	**Gender**	**Certificate**	
D_1 (T1)	f	W	m	K	Complex decisions
D_1 (T2)	f	K	m	W	
D_2	f	K	m	K	Gender decisions
D_3	f	W	m	W	
D_4	f	W	f	K	Certificate decisions
D_5	m	W	m	K	
D_6	f	K	m	–	Simple decisions
D_7	f	W	m	–	
D_8	f	–	m	K	
D_9	f	–	m	W	

Notes: (i) "D_1", "D_2", etc. refer to decision 1, decision 2, etc., "T1" and "T2" refer to TREATMENT 1 and 2, "f" refers to female candidate, "m" refers to male candidate, "W" refers to the word task, while "K" refers to the general knowledge task. (ii) 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.

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

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 (employers knew this). 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] 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 decision 1 (the complex decision) and an additional 6€ if their candidate choice in a randomly drawn decision from 2 to 9 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 to 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 as a strict one, which will be especially important when we consider the hiring decisions that are explicitly about gender only (i.e., the gender decisions).

The order of decisions was partially randomized at the individual level. The treatment-specific version of decision 1 was always taken first. It was followed by a block with decisions 2–5, randomly ordered, and then a block with decisions 6–9, randomly ordered. This 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 decision 1, we also measured the employers' beliefs about the informativeness of each of the

[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.

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

two certificates about performance in the job task.[15] The reason for doing this was to assess whether participants shifted their beliefs in order to justify their decision 1 hiring choices as being purely qualification-based. For example, an employer holding a gender bias in favor of the male candidate, who happened to have a word certificate in decision 1 (i.e., who was presented in randomized treatment 2), might adjust their belief about the informativeness of the word certificate for job performance upwards to justify choosing the male candidate.

2.2 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.

[15] 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).

2.3 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 N = 119 (49.6%) were female. The average age was 24.5 years (SD: 4.9 years) and the majority were students in a STEM (50%) or Economics/Business program (31%). 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 were female. The sessions were conducted between October 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 took place within session, resulting in $N_1 = 119$ employers in Treatment 1 and $N_2 = 121$ in Treatment 2. Demographics by treatment assignment are reported in Table 5 of Appendix B.

3 Aggregate Analysis: Gender Bias and Discrimination

As we are not imposing any assumptions on subjective beliefs or assuming a particular model of behavior, it is worthwhile clarifying what we mean by gender discrimination in our setting. We refer to the empirical/behavioral phenomenon that an employer's hiring decision is causally affected by the candidates' gender. Our design allows us to identify and measure such discrimination by presenting employers with several "symmetric-in-gender" pairs of binary decisions, in which the only difference is that the gender of the two candidates is reversed between CVs; e.g., the two complex decisions, D_1, where one treatment has (m,K) vs. (f,W), and the other has (f,K) vs. (m,W) (see Table 1). Any significant difference between the rates at which a given certificate is hired between two such gender-symmetric decisions can be cleanly attributed to the effect of gender information (the "gender of the certificate" has a causal effect on its hiring rate). Equivalently, if employers' choices were all about qualifications only, the rate at which men are hired in one such decision would equal the rate at which women are hired in the other, so both genders are hired at the same rate of 50% overall, when aggregating the two decisions. By measuring the difference in this overall rate between the two genders, which we will call *gender bias*, we can therefore identify whether some employers discriminate; e.g., if men are hired at a significantly higher rate, there is a gender bias against women, implying that some employers hire men over women irrespective of the two qualifications featured in the two decisions. Moreover, this aggregate-level measure of gender bias provides us with a lower bound on the amount of discrimination (e.g., against women) that occurs at the individual level.

We formally illustrate this relationship between aggregate level gender bias and individual level discrimination for the complex decisions (D_1) using Table 2. The row indicates the

Table 2: Identification of gender bias and discrimination.

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

preferred candidate in D_1 of treatment 1, (m,K) vs. (f,W), and the column indicates the preferred candidate in D_1 of treatment 2, (f,K) vs. (m,W). The cells indicate the proportions of employers for all four possible preference combinations, i.e., the joint distribution over preference types, so 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).[16]

Let x be the fractions of employers who would choose the candidate with the K certificate in the "row decision" (where this candidate is male, treatment 1) and similarly let $(1-x')$ denote the fraction who would choose the candidate with the K certificate in the "column decision" (where this candidate is female, treatment 2). 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 $b < 0$.[17] 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

[16]Note, since each participant in the experiment was either presented with the decision D_1 from treatment 1 *or* from treatment 2 (in order to avoid making the comparison salient), we do not directly observe the individual types represented in Table 2.

[17]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 ($\sigma_m = \sigma_f = 1/2$).

that $\sigma_m \geq 0$, we obtain $\sigma_f \geq -b$. (Note, one such lower bound is positive if and only if the other is negative.) Of course, x and x' immediately yield also upper bounds, namely $\sigma_m \leq \min\{x, x'\}$ and $\sigma_f \leq \min\{1-x, 1-x'\} = 1 - \max\{x, x'\}$.

The simple decisions, D_6 and D_8 (knowledge certificate vs. no certificate) as well as D_7 and D_9 (word certificate vs. no certificate), also constitute gender-symmetric pairs of decisions for identifying discrimination via gender bias as with the two versions of D_1. Hence, in the sections below, we will carry out the very same analysis of discrimination using our measure of aggregate gender bias, allowing us to also compare discrimination across the different kinds of decisions. We will additionally relate our findings to the "pure" gender bias observed in D_2 and D_3, where candidates have identical qualifications and differ solely in their gender, so that they are explicitly about gender. Since all decisions other than D_1 are made by all employers (in both treatments), we are able to compute gender bias using the pooled/full sample for all of these analyses.[18] However, we will also summarize the two treatments separately and compare them at the end of this section.

In the analyses below, proportions are compared using χ^2 tests of independence.[19] That is, we also use such tests to evaluate the null hypothesis that $b = 0$ against the two-sided alternative $b \neq 0$, in line with the definitions above (which is equivalent to a χ^2 goodness-of-fit test of the null hypothesis that men are hired with a propensity of $\overline{x} = 0.5$). The reported 95% confidence intervals for proportions are Wilson score intervals.

3.1 Gender bias in simple and complex decisions

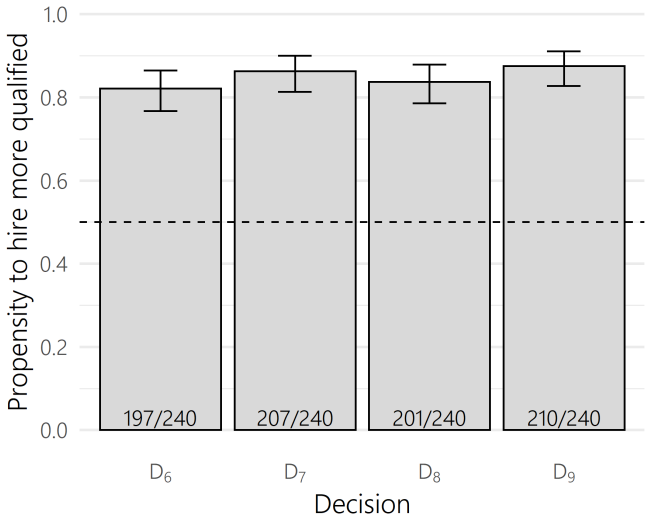
It will be useful to first consider the simple hiring decisions, all of which are between one candidate with a certificate (knowledge or word) and another without a certificate. The two candidates being compared always differ in their gender in these simple hiring decisions. We shall refer to the candidate with the certificate as being *more qualified*. Figure 2 reports the average propensity to choose the more qualified candidate for each of these four decisions. The two left-most bars show the hiring propensities when the female candidate has a knowledge or word certificate, and the competing male candidate has no certificate. Similarly, the two right-most bars show the hiring propensities when the more qualified candidate is male.

[18] It is computed exactly as in equation (1) but where x and x' both come from the full sample. Our aggregate analysis thus ignores within-subjects information, which will be key in our individual analysis in Section 4 below.

[19] Note that as soon as we pool data from decisions within treatments, inference based on the assumption of independence of observations is strictly speaking not appropriate, since the same participants contribute multiple observations. This can be taken into account via using the Cochran-Mantel-Haenszel (CMH) test statistic instead of the standard Pearson χ^2 statistic (Agresti, 2013). Intuitively speaking, such tests are based on the null hypothesis that the the number of employers switching in one direction between decisions is equal to the number of employers switching in the other direction, thereby accounting for within-subject correlation of responses. Therefore, in the rare cases where our analysis pools multiple decisions of the same participants, we employ the CMH test.

In all four decisions, the vast majority of hiring choices favor the more qualified candidate (between 80% and 90%), irrespective of whether that candidate is a man or a woman. Therefore, we neither observe gender bias in hiring for these simple decisions when pooling decisions 6 and 8 (where candidates either have a knowledge certificate or none; $\chi_1^2 = 0.35$, p = 0.56), nor when pooling decisions 7 and 9 (where candidates either have a word certificate or none; $\chi_1^2 = 0.24$, p = 0.62). The fact that approximately 15% hire the candidate without a certificate is consistent with non-discrimination and heterogeneous beliefs.[20]

Figure 2: Initial hiring choices in simple decisions.



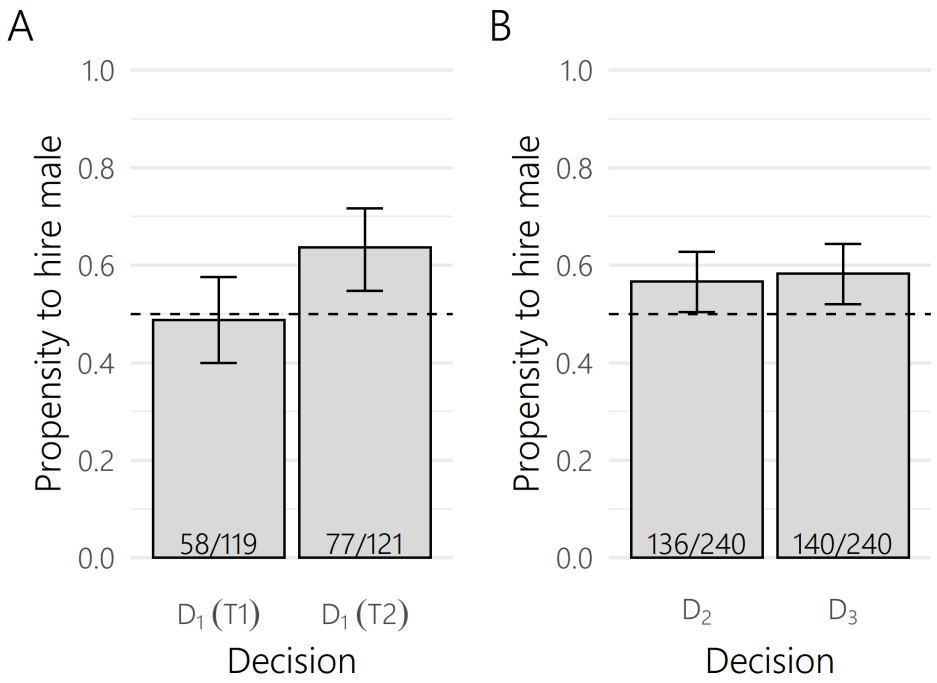
Notes: (i) The figure shows the propensity to hire the more qualified candidate in the Decisions 6-9, where the gender and certificate of the more qualified candidate varies, (ii) As described in Table 1, in D_6 and D_7 the more qualified candidate is a female, with a knowledge certificate (f, K) in D_6, and a word certificate in D_7 (f, W). In D_8 and D_9, the more qualified candidate is male, with D_8 corresponding to (m, K) and D_9 to (m, W), (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.

Turning to complex hiring decisions, Panel A of Figure 3 reports the propensity to hire the male instead of the female candidate when the two candidates hold different certificates. Since they are *differently qualified*, it is more opaque which candidate is *better qualified*. The left bar shows that when the male candidate has the knowledge certificate and the female candidate has the word certificate (treatment 1), there is no significant difference between the rate at which male and female candidates are chosen ($\chi_1^2 = 0.08$, p = 0.78), with the

[20]If we were to impose the assumption that every employer "must" perceive either certificate to be good news about job performance, so that the candidate with the certificate is unambiguously *better qualified*, then employers' choices of candidates without a certificate would constitute a significant amount of individual level gender discrimination. However, even then, it is worth noting that it affects men and women to the same extent and cancels out in the aggregate. Therefore, any individual level gender discrimination in these decisions does not translate into aggregate gender bias.

male candidate chosen 48.7% of the time. However, when these qualifications are reversed (treatment 2), the male candidate is chosen 63.6% of the time, which implies that male candidates are chosen substantially more often than female candidates ($\chi_1^2 = 9.00$, p = 0.0027). Averaging across these two scenarios, men are hired 56.2% of the time, which is significantly higher than the gender-neutral benchmark of 50% ($\chi_1^2 = 3.68$, p = 0.055). This corresponds to a gender bias of 12.4% against women. It shows that in the more subjective, complex hiring decisions, we observe male-favoring discrimination amongst the same group of employers who display no gender bias in the simple hiring decisions, with a lower bound of 12.4% on the proportion of employers that discriminate against women.

Figure 3: Initial hiring choices in complex (A) and gender (B) decisions.



Notes: (i) Panel A reports the propensity to hire the male candidate in the Decision 1 between the two treatment conditions. These decisions involve a male and a female candidate with different qualifications, (ii) Panel B reports the propensity to hire the male candidate in Decisions 2 and 3, where the employer chooses between a male and female candidate with identical qualifications, (iii) Error bars show 95% confidence intervals, (iv) The dashed horizontal line indicates an equal aggregate propensity to hire both candidates in a particular choice.

3.2 Gender bias in gender decisions

The discussion above has illustrated that: (i) when one candidate is *more qualified*, employers do not display a gender bias in their hiring decisions, and (ii) when candidates are *differently qualified* and the qualifications of the two candidates cannot be unambiguously ranked, the same group of employers preferentially hire men.

This raises the question of whether employers display a gender bias when candidates are *equally qualified*. Here, we therefore examine the extent to which a gender bias arises even in decisions between identically qualified candidates that differ only in their gender, i.e., in our so-called gender decisions (D_2 and D_3). Given identical qualifications, at an individual level, non-discrimination requires indifference. At an aggregate level, the absence of a gender bias requires that men and women are hired equally often.[21] Panel B of Figure 3 displays the propensity of employers to choose the male candidate. When both candidates have a knowledge certificate, men are hired in 56.7% of cases, and when both candidates have a word certificate, men are hired in 58.3% of cases. Both proportions are significantly different from 50% ($\chi_1^2 = 4.27$, p = 0.039, and $\chi_1^2 = 6.67$, p = 0.0098, respectively), and together imply an average gender bias of 15%. Thus, even when CVs are otherwise identical and it is salient that a choice is related directly to the candidate's gender, there is a significant gender bias against women in hiring and the magnitude is similar to that in the complex hiring decisions.

3.3 Summary of aggregate behavior in all decisions (by treatment)

Figure 4 provides an aggregate-level summary of the choices made in all decisions, by treatment. In Figure 4, each edge denotes a binary comparison between two nodes. The nodes each describe a particular candidate profile. For example, the north-east node in each panel refers to *(f,K)*, which reflects a female candidate with a knowledge certificate. When compared to another node, an arrow pointing away from a node reports the propensity to choose the profile at the base of the arrow. Similarly, the parallel arrow pointing towards a node shows the propensity to choose the comparison profile.

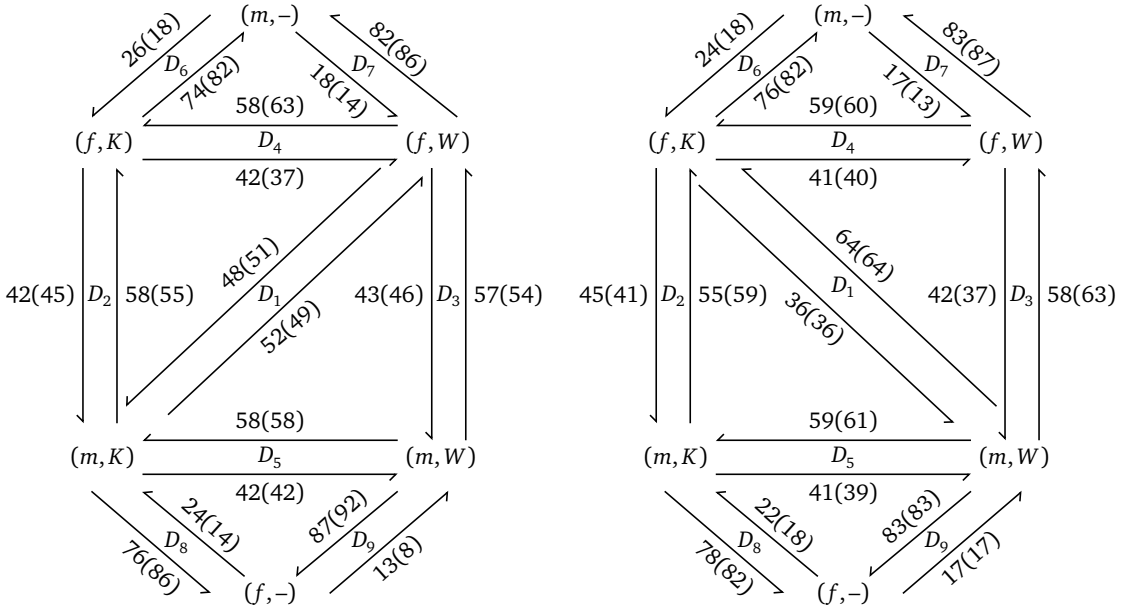
To interpret the figure, it is important to recall that after making each initial hiring decision, employers could "sell" their choice for 0.10€. So far, we have focused exclusively on studying these initial hiring choices in our aggregate analysis. Figure 4, however, summarizes the propensity to hire each candidate both in these initial hiring decisions as well as in the final decisions that after the selling option is exercised. We refer to these final decisions as "ultimate hiring choices". To calculate these ultimate hiring propensities, we make use of the feature of our design that whenever an employer sells their initial decision, there is an equal probability of each of the candidates being hired. Therefore, sold choices are assigned with equal probability to each candidate.

In Figure 4, the initial decision propensities (in %) are reported in parentheses next to the ultimate decision propensities. Each arrow in a figure has an associated initial and ultimate propensity, and the propensities associated with pairs of parallel arrows always sum to 100. For example, the figure shows that profile *(f,K)* is involved in decisions, D_2, D_4,

[21] In relation to the simple framework developed in Section 3 above, we can measure gender bias directly from each single decision D_2 or D_3; formally, we simply replace the average $\overline{x}$ in equation (1) with the fraction of employers choosing the male candidate in that decision.

and D_6 and provides a full summary of all of these decisions for both treatments. In decision D_6 in treatment 1 (left panel), *(f,K)* was chosen over *(m,-)* in 82% of initial hiring choices and in 74% of ultimate hiring choices.

Figure 4: Ultimate (initial) choice propensities in both treatments



Notes: (i) The left-hand panel summarizes the aggregate choices from Treatment 1 ($N_1 = 119$), while the right-hand panel describes the same information for Treatment 2 ($N_2 = 121$), (ii) Each edge of the figure contains a pair of parallel arrows and corresponds to a binary comparison, D_i $(i = 1, \ldots, 9)$, between two candidates (g, C) and (g', C'), which are described at the nodes of the figure. Each arrow is associated with two propensities, $X(Y)$, associated with the initial and ultimate choices 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.

Since the treatments differ only in terms of which version of D_1 employers make, whereas decisions D_2 to D_9 are made by all employers, we can directly compare these eight decisions between treatments. Figure 4 shows clearly that the aggregate choice propensities are virtually identical between the two treatments for all eight decisions that they have in common (i.e., excluding the complex decisions, D_1). This is true for both the initial and ultimate hiring choices. Importantly, the observed gender bias remains very similar after accounting for sold choices.

In addition to providing further evidence on the simple, complex and gender choices discussed above, Figure 4 also reveals how employers value the information content of the two qualifications. This is done by examining the certificate decisions (D_4 and D_5), where there is no gender bias by construction because the two candidates are of the same gender and differ solely in their certificates (word vs. knowledge). In both decisions, a majority of close to 60% of employers hire the candidate with the word certificate, suggesting that this certificate is typically considered to be more informative about good job performance (irrespective of the gender of the pair of candidates).

To summarize: In simple hiring decisions, where one candidate is more qualified, employers show no aggregate gender bias in hiring. However, as soon as the pair of candidates' qualifications no longer display such a natural ranking—whether it is because the qualifications are actually identical or because their ranking is a matter of subjective assessment—a significant gender bias against women emerges. Comparing these two cases, i.e., gender decisions and complex decisions, we observe that the bias is of a similar magnitude, with men hired at a rate of approximately 57% and women 43%. While in the gender decisions, this hiring bias is largely independent of the specific certificates that the pair of candidates hold, in the complex decisions, the certificates do play a role: There, if the woman holds the certificate that is considered to be the better certificate by the majority of employers, namely the word certificate, this offsets the pure gender bias, such that in this scenario no gender bias is observed overall. However, when the man holds the certificate that is perceived to be better, the gender bias is exacerbated, and to a similar extent. This implies that on average, across the two scenarios, the gender bias is of a similar magnitude to that observed in the gender decisions.

3.4 Motivated reasoning through means of "redefining merit"

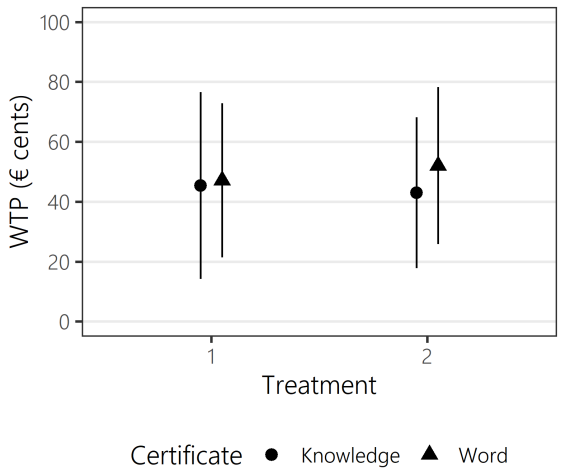
In this section, we explore the idea that hiring decisions between candidates that are *differently qualified* presents employers with an opportunity to hide or obscure their discrimination from themselves and others. An employer who holds beliefs that are biased against a particular gender faces a very different decision problem in the complex decisions in comparison to the gender decisions. In the case of the complex decisions, the heterogeneity of the qualifications of the two candidates provides the employer with the opportunity to make a gender-biased decision, while convincing themselves that in fact their decision is based solely on the different qualifications of the candidates. 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.

This phenomenon has also been studied by psychologists who refer to it as the propensity to "redefine merit" in a motivated manner. For example, Uhlmann and Cohen (2005) describe employers shifting their preferences over the value of different qualifications as a way to "allow people to maintain an image of themselves as objective and principled" (p. 479) despite discriminating by gender when hiring employees.[22] We investigate this psychological channel by comparing employers' beliefs about the predictive value of each of the two certificates (knowledge, word) for the job task between the two treatments. These beliefs are elicited immediately after their complex decision, D_1. Since the treatments differ

[22]One potential foundation for this idea is that image concerns are driven by a signaling motive, with individuals trying to maintain positive image of "who they are" through identity management (Bénabou and Tirole, 2011). In the closely related social psychology literature on aversive racism, signaling motives are also discussed as playing a key role (Hodson et al., 2010).

in D_1 in terms of whether the male of female candidate holds the word certificate (and vice versa), an employer who makes a gender-biased decision in D_1 might (subconsciously) justify it to themselves by inflating the importance that they assign to the certificate that their preferred candidate holds.

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 (word) certificate in comparison to a randomly drawn candidate, (ii) Sample means with ± one sample standard deviation are reported.

Figure 5 reports the employers' willingness to pay (WTP) for a candidate with a specific certificate in comparison to a randomly drawn candidate in each treatment (see footnote 15 for details). Visually, this shows that the mean WTP for the two certificates in treatment 1 are close to one another, while in treatment 2, the word certificate (which is held by the male candidate) has a slightly higher mean WTP. We test whether there is a statistically significant difference between the beliefs held by employers regarding the value of the two certificates between the treatments by estimating a difference-in-difference model. The point estimate of the interaction coefficient is 7.3 € cents, indicating that employers have a slightly higher WTP for the word certificate relative to the knowledge certificate in the treatment where the male candidate holds the word certificate, however this difference is not statistically significant ($t_{224} = 1.48$, p = 0.14). 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 relative value of the two certificates, we can use these to impute the missing data for those participants who switched multiple times in the price lists (This occurred in 42 out of 480 responses, of which 23[19] are in treatment

1[2], and 20[22] are for the knowledge [word] certificate).[23] This exercise reveals a similar point estimate for the effect size, namely that employers' WTP is 8.1€cents higher for the word certificate when it is held by the male candidate. However, here this coefficient is more precisely estimated—perhaps due to the larger sample size achieved when we are able to include the participants who switched multiple times in the price list ($t_{239} = 1.77$, p = 0.078; Appendix Table 7, columns 1, 2).

It is worthwhile noting that we would expect a causal effect only for those employers that have an interest in covering up their discriminatory behavior, which may explain the rather small estimated effect size. The following individual analysis will allow us to shed further light on this.

4 Individual Analysis: Explicit and Implicit Discrimination

Psychological research on discrimination emphasizes the potential conflict between an individual's discriminatory beliefs or tastes and social norms that portray discrimination as undesirable (see, e.g., Snyder et al., 1979; Darley and Gross, 1983; Greenwald and Banaji, 1995; Dovidio and Gaertner, 2004; Greenwald and Krieger, 2006).[24] For instance, employers who believe that men are better at doing a particular type of job but also do not want to discriminate (or to appear to discriminate) against women may experience such a conflict when making hiring decisions. The key idea here is that the way in which this tension will be resolved may hinge on how easily the decision can be attributed to discrimination. Specifically, whether such an employer will end up discriminating depends on the degree to which their individual choice can be objectively identified as a violation of a social norm: The employer will tend to comply with the norm in decisions where violations are easy to detect ("explicit violations") and deviate from the norm where violations are hard to detect ("implicit violations").

[23] To do this, we exploit the strong theoretical and empirical correlation between the responses in the price list-based belief elicitation 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 9 in Appendix B). Specifically, for the imputation, we predict the missing values in WTP by using the estimates from an OLS regression of the two measures. We then impute the WTP measure from the unintentivized measure for missing observations separately per certificate. We round them to the nearest 10€cents so that they take a similar form to the non-missing observations measured directly from the price lists.

[24] One reason for this is that discrimination as a topic fits neatly into the wave of work in psychology focused on studying scenarios where there is a tension between explicit and implicit attitudes. As noted by Greenwald and Krieger (2006), implicit attitudes are typically most interesting when they come into tension with explicit attitudes and this is often the case for discriminatory views. This literature has often used the implicit association test (IAT) to try to tease apart implicit and explicit attitudes (Greenwald et al., 1998).

4.1 Classifying discrimination types

We now turn to the within-dimension of our experiment, and how it allows us to operationalize these notions and identify explicit and implicit discrimination. Essentially, we first use the two gender decisions (D_2 and D_3) to classify individual employers into explicit discrimination types, based on the *individual* gender bias their choices explicitly reveal here, and then separately measure the gender bias of these types in the complex decisions (between-subjects). Implicit discrimination against women is then identified as discrimination against women by employers that do not explicitly discriminate against women; analogously, for implicit discrimination against men. Observe here that the complex decisions involve the same qualifications as the gender decisions (candidates have either a knowledge or a word certificate), but such that the two candidates differ in terms of both gender and qualification, thus providing scope for employers to rationalize discriminatory decisions as being non-discriminatory and obscuring discrimination.

Any employer *strictly* choosing one candidate over the other, by not selling the initial choice, *in a gender decision* reveals both (i) that they believe gender matters here, and (ii) that they are willing to let this information affect their hiring decision. Accordingly, we interpret their hiring choices in these decisions as overt expressions regarding gender discrimination: We say that an employer *discriminates explicitly* if in at least one of the two gender decisions she does not sell her initial choice.[25] If she does sell both initial choices, we say that she *explicitly does not discriminate*. We will further distinguish explicit discriminators according to their individual gender bias and say that an employer *explicitly discriminates against men (resp., women)* if she explicitly discriminates while always choosing the woman (resp., man) initially. Here, "always" refers to making a consistent gender-choice in both gender decisions. Explicit discriminators that cannot be categorized in this way, because they choose the man in one decision but the woman in the other, will be referred to as "mixed" explicit discriminators.

Relating this classification to social norms, those employers that explicitly do not discriminate (and only these) comply with a strong gender-symmetric norm of non-discrimination that prescribes fully gender-blind decision making. By contrast, the norm of not discriminating *against women* is satisfied also by those employers that explicitly discriminate against men. Though weaker and gender-asymmetric, this norm allows for "affirmative action favoring women," in particular a *ceteris paribus* rule for preferential hiring of women over equally qualified men, as some organizations explicitly feature in their job adverts (e.g., in Germany). Our type classification separately considers these two types of explicit norm compliance.

Due to the fact that we do not observe aggregate-level treatment differences in the gen-

[25] Recall that after making an initial "forced" hiring choice between two candidates, employers were given the option of having this choice replaced by a random draw and receiving a sure extra payment of €0.10.

der decisions (D_2 and D_3), we can carry out the same between-subjects analysis as in the previous section separately for each type of employer, by computing gender bias as in equation (1). Thus, our methodology operationalizes the formal framework to behaviorally identify implicit preferences recently proposed by Cunningham and de Quidt (2022), whilst refining the explicit type classification by distinguishing two different candidate norms. Table 3 reports the distribution of these discrimination types for each treatment. It also reports the propensity of each type to hire the male candidate in the complex decisions.

Table 3: Propensity to initially hire the male candidate in D_1 by discrimination type.

Discrimination type	**Frequency**		**Propensity m in D_1**	
	Treatment 1 (1a)	Treatment 2 (1b)	Treatment 1 (2a)	Treatment 2 (2b)
Explicit discrimination	68.1	72.7	58.0	63.6
– against men	13.4	12.4	62.5	66.7
– against women	26.1	25.6	61.3	67.7
– mixed	28.6	34.7	52.9	59.5
Explicit non-discrimination	31.9	27.3	28.9	63.6

Notes: (i) Columns (1a) and (1b) report the frequency distribution of different discrimination types (in %), based on the classification using D_2 and D_3, (ii) Columns (2a) and (2b) report the propensity to initially hire the male candidate in D_1, by discrimination type within each treatment group (in %).

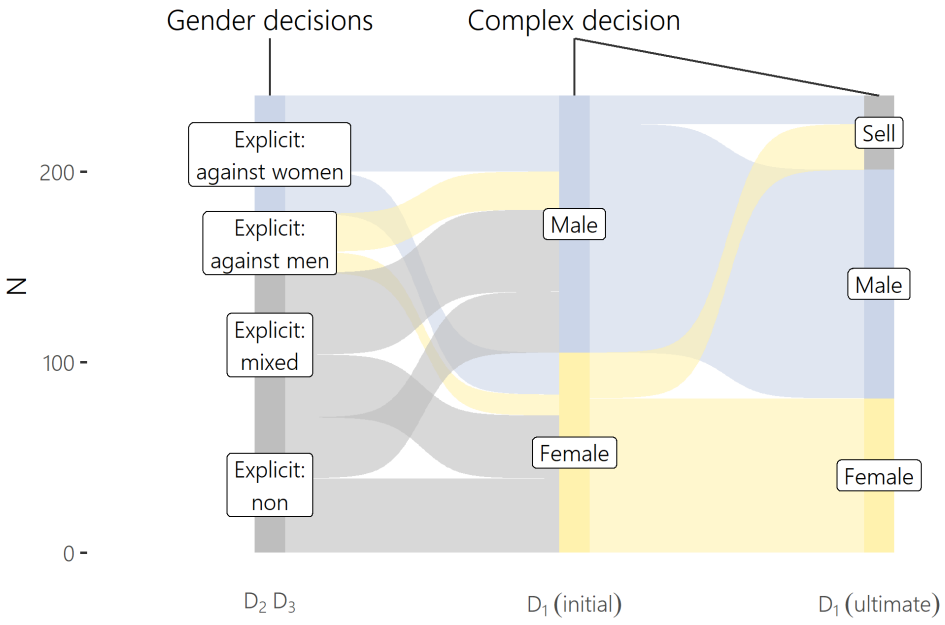
In line with the earlier observation that there is no treatment difference in the gender decisions, the type distributions in our two treatments are very similar. Approximately 70% of the employers discriminate explicitly, and they are roughly twice as likely to do so against women as against men. Most remarkably, however, employers that discriminate explicitly against *men* exhibit substantially male-biased hiring propensities in the complex decisions D_1. Moreover, comparing columns (2a) and (2b) shows that their propensities to hire the male candidate is strikingly similar across both decisions/treatments to those of employers that explicitly discriminate against women. Accordingly, the gender bias against women of these two types in D_1 is similar: We thus obtain nearly identical lower bounds on the share of employers discriminating against women—hiring the male candidate irrespective of qualifications—of 29.2% and 29.0%, respectively, as well as also finding very similar upper bounds, where the shares of these two groups discriminating against women in D_1 may be as high as 62.5% and 61.3%, respectively.[26] While these calculations are for initial hiring choices only, hardly any of these employers sell their initial choices in D_1, and we observe similar hiring propensities and gender bias estimates in the ultimate hiring decisions.[27]

[26]Recall from Section 3 above that we can calculate the lower bound on σ_m using $\sigma_m \geq b = x - (1 - x')$. Here, this yields $b = 62.5 - (100 - 66.7) = 29.2$ for those who discriminate explicitly against men, and $b = 61.3 - (100 - 67.7) = 29$ for those who discriminate explicitly against women. The upper bound is given by $\sigma_m \leq \min\{x, x'\} = \min\{62.5, 66.7\}$ for those who discriminate explicitly against men, and $\sigma_m \leq \min\{61.3, 67.7\}$ for those who discriminate explicitly against women.

[27]The propensities in Treatment 2 are identical when accounting for sold choices in comparison to the initial choices; in treatment 1, there is a slight drop from 62.5% to 59.4% among employers that explicitly discriminate against men and a slight increase from 61.3% to 62.9% among employers that explicitly discriminate against women.

In contrast, explicit non-discriminators display only a small gender bias of 7.4% in the complex decisions, and this goes in the opposite direction, namely against men. However, once we account for sold hiring choices, this gender bias against men is reversed and we observe a small gender bias against women of 2.9%.[28] Taken together, this evidence shows that explicit discriminators account for essentially all of the gender bias against women in the complex decisions.

Figure 6: 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 (D_2 and D_3), and their decision making in the complex decision (D_1). (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 in the experiment (240).

Figure 6 yields further insight into the relationship between employer types and implicit discrimination by providing an illustration of how employer decision making in the gender decisions (D_2 and D_3) relates to their hiring choices in the complex decisions (D_1). The figure contains three columns, with the left column displaying the fraction of each employer type in the gender decisions, the middle column showing the fraction choosing the male and female candidate initially in the complex decision and the right column reporting the fraction choosing the male candidate, female candidate or selling their choice in their ultimate decision. The key feature of the figure is that it also shows the flows between these columns, in relation to the entire sample of all 240 employers. This illustration reveals several interesting insights: (i) male candidates tend to be favored over female candidates overall, (ii)

[28]This is because these employers sell their initial choices of the female candidate at a higher rate in treatment 1: Male hiring then increases from 28.9% to 40.8%.

employers who are classified as discriminating explicitly against women and those who are classified as discriminating explicitly against men are *both* more likely to choose the male candidate in the complex decisions, (iii) only a small fraction of initial D_1 decisions are sold, but the propensity to sell initial hiring decisions where the female candidate was chosen is higher, (iv) the vast majority of the employers who hired the female candidate in the complex decision are classified as either non-discriminators or mixed explicit discriminators in the gender decisions. Together, these empirical patterns demonstrate a strong asymmetry in the treatment of male and female candidates by employers.

4.2 Implicit discrimination and heterogeneity by discrimination types

How do these findings relate to implicit discrimination? First, the large bias against women that we observe in the complex decisions among employers that discriminate explicitly against men identifies a strict notion of implicit discrimination. It means that there are employers that explicitly comply with a norm of not discriminating against women, in the gender decisions, yet do discriminate against women implicitly, in the complex decisions, where they would choose the male candidate regardless of qualifications. Their choices correspond to the "figure-8" pattern that Cunningham and de Quidt (2022) highlight as the key distinctive choice implication of their formal model of implicit preferences, which cannot be rationalized by any single transitive preference over alternatives. In the "figure-8" pattern, explicit preferences regarding a particular attribute (gender) that are revealed when only this attribute differs (gender decisions, equally qualified candidates) are essentially the opposite of those that are revealed when this difference is mixed with difference on another attribute (complex decisions, differently qualified candidates), hence implicit.

Turning more specifically to our setting, a significant share of employers that reveal an explicit bias against men via the gender decisions are revealed to hold an implicit bias against women via the complex decisions. This cannot be rationalized by any single transitive preference, in particular by expected utility maximization, even allowing for any subjective beliefs.[29] Moreover, it is striking to note that, in addition to the male gender bias in the two complex decisions, even the hiring propensities in both of these decisions are basically identical between employers that explicitly discriminate against men (in favor of women) and employers that discriminate explicitly against women. This is consistent with an interpretation where the only difference between these two types is whether or not they

[29]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 D_1 (T2) and D_2). 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.

are concerned about explicitly discriminating against women and thus violating an anti-discriminatory norm here.

Second, the absence of any significant gender bias in the complex decisions among employers that explicitly do not discriminate suggests that those who explicitly adhere to the strong gender-symmetric norm of not discriminating at all, by contrast, comply with non-discrimination more generally, i.e., consistently choose between qualifications (the knowledge and word certificates) regardless of gender across all of decisions 1, 4 and 5. Indeed, we find that the majority of these explicitly non-discriminating employers (55.3% in Treatment 1, and 54.5% in Treatment 2) make choices that either (i) consistently favor the candidate with a specific certificate in all three hiring decisions where one of the two candidates holds the knowledge certificate and the other holds the word certificate, irrespective of the candidates' gender (i.e., D_1, D_4 and D_5), or (ii) consistently express (near-) indifference.[30] These "non-discriminators" make up around 16% of all employers, whereas the probability of observing such non-discrimination across the five decisions under random choice is less than 4%.[31] Notably, even the other explicit non-discriminators, who are inconsistent in their choice of which certificate to favor across D_1, D_4 and D_5, do not exhibit any significant gender bias in the complex decisions (average initial and ultimate hiring rates of men are 48.0% and 52.1%, respectively), suggesting that these are best thought of as noisy non-discriminators.

Finally, we revisit discrimination in the simple decisions as well as the observed treatment effect on beliefs regarding the value of certificates, and examine heterogeneity based on our type classification. First, we find that employers who discriminate explicitly against a given gender (based on the gender decisions) also display a bias against the same gender in the simple decisions: Pooling employers of a given type across treatments, and aggregating over all four simple decisions, the rates at which men are hired are 46.8% for employers that discriminate explicitly against men and 57.3% for those that discriminate explicitly against women ($\chi^2_1 = 4.74$, p $= 0.029$, for the comparison). Consequently, when ranking candidates' qualifications is less subjective, those employers that explicitly discriminate against men still tend to do so, and in any case do not discriminate against women, in contrast to their implicit discrimination in the complex decisions. Second, we find that it is also exactly these two types of employers whose valuations of the two certificates are subject to signifi-

[30] This means one of the following three patterns: Always hire the K candidate, always hire the W candidate, or always sell. Note that even when just considering the two certificate decisions (D_4 and D_5), choosing the K candidate in one and the W candidate in the other constitutes gender discrimination; the dismissed candidate is then dismissed because of *both* candidates' particular gender. As an illustrative example, it is akin to making a hiring decision between two women for a particular job and choosing the candidate who is more caring, but then when making the exact same hiring decision between two men for the same job, choosing the candidate with better numeracy skills.

[31] Supposing an employer always picks at random, both initially and then also in deciding whether to sell (two binary decisions), this probability is the product of probability 1/4 of selling in both gender decisions and probability 2/64+1/8=5/32 of consistently choosing between the two certificates in the complex as well as the two certificate decisions.

cant treatment effects, thus both ex post redefining merit depending on which certificate the male candidate had in the preceding complex decision (see Appendix Table 7, columns 2a–d). In particular, even those employers that explicitly discriminate against women appear to justify their predominantly male hiring as qualification-based.[32] This lends additional support to the basic idea that discrimination is subject to a tension between social norms against gender discrimination (in particular discrimination against women) and beliefs that gender matters (in particular gender stereotypes).

In summary, the significant aggregate gender bias against women in the complex decisions is largely driven by employers that discriminate explicitly, a significant share of whom are revealed to hold an implicit bias against women that contradicts their explicit bias against men. Indeed, this group's biased hiring behavior in complex decisions is indistinguishable from that of employers that explicitly discriminate against women (64.5% initial hires of the man for both types, see also Figure 6). Moreover, both of these employer types have a tendency to adjust their assessment of qualifications to favor the qualification held by the male candidate and thus align their discriminatory behavior in complex decisions with a norm of not discriminating against women.

5 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. (2020) 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.

[32]Explicit discriminators against men attach the same value to both certificates in treatment 1 (55€ cents), but a significantly lower value to each of them in treatment 2; while valuations of the two certificates in treatment 2 exhibit quite a gap (34€ cents for K vs. 45€ cents for W), the corresponding interaction coefficient is not statistically significant. In contrast, for employers that discriminate explicitly against women, this interaction term is significant.

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.[33]

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%).[34] Examining actual performance differences also allows us to gauge to what extent such non-discrimination is really "irrational" here.

5.1 Do men actually perform better than women? Stereotypes and aggregate gender bias

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 7, and Table 8 of the Appendix B adds standard descriptive 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_{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 4, 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.

In contrast, when we look at the actual decisions of our employers by aggregating over all decisions, we find the opposite, namely a bias towards men (the only decisions where women are hired more often are those simple ones where they were more qualified). 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

[33] 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.

[34] In addition to explicitly not discriminating in D_2 and D_3, being consistent with non-discrimination across the nine decisions requires choosing consistently between qualifications K, W and – in all other decisions. Under random choice, the probability of satisfying this standard is less than 0.5%.

conclusion, by drawing on the 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 statistically greater than 50 ($t_{79} = 2.99$, p = 0.0037).[35]

Figure 7: 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. Therefore, 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.

5.2 Statistical accuracy of discrimination and non-discrimination

Table 4 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

[35]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 ($t_{79} = 3.98$, p < 0.001). Both distributions are visualised in Appendix B.

as the random tie-breaking in determining employers' payoffs). It tells us who should be chosen in each of the nine 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 7, where despite being more qualified a randomly drawn woman in our sample is not more 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 2's complex decision (D_1), 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.

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 4 shows in rows 11–13, when ignoring gender information, statistically, the knowledge certificate should be favored over the word certificate, and either of the two indicates better performance than that of a purely random candidate (corresponding to $(-,-)$ in the table). Yet, a majority of around 64% of those employers that are "non-discriminators" across of decisions 1–5, which involve these two certificates, hire the word over the knowledge certificate in the two certificate decisions, which are only about qualification (64.3% and 63.9%in Treatment 1 and 2, respectively, and 61.9% and 63.9% in terms of ultimate as opposed to initial hiring). In this sense, also the majority of gender non-discriminators we observe could be classified as statistically inaccurate as well.

When it comes to reducing discrimination via information campaigns, as advanced by Bohren et al. (2020), 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 their beliefs are corrected. For example, if qualifications that men are more likely

Table 4: True probabilities of which candidate is the better one in the job task, with random tie breaking.

	A	B	Pr(A better)	Pr(B better)	Pr(tie)
D_1 (T1)	(f,W)	(m,K)	0.426	0.574	0.121
D_1 (T2)	(f,K)	(m,W)	0.499	0.501	0.169
D_2	(f,K)	(m,K)	0.554	0.446	0.171
D_3	(f,W)	(m,W)	0.374	0.626	0.107
D_4	(f,W)	(f,K)	0.356	0.644	0.313
D_5	(m,W)	(m,K)	0.540	0.460	0.188
D_6	(f,K)	$(m,-)$	0.633	0.367	0.164
D_7	(f,W)	$(m,-)$	0.500	0.500	0.139
D_8	$(f,-)$	(m,K)	0.456	0.544	0.148
D_9	$(f,-)$	(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

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.

6 Conclusion

The economics literature has typically dichotomized discrimination into taste-based (Becker, 1957) and (accurate) statistical discrimination (Phelps, 1972; Arrow, 1973). Even though our experimental design (i) does not permit employers to preferentially reward or choose to interact with candidates from one gender, and (ii) involves a job task where there is no gender gap in performance on average we still observe substantial discrimination against women in the hiring task. In particular, we observe evidence of both explicit and implicit belief-based discrimination, with women discriminated against more often than men. Further, our type classification exercise showed that when hiring decisions are revealing about an individual's gender bias (explicit discrimination), we observe one group of employers who discriminate against women and another group who discriminate against men. However, strikingly, both these groups display a similarly substantial gender bias against women in the complex decisions (implicit discrimination), where attribution to gender information is obscured. Taken together, our results are consistent with the idea that a majority of em-

ployers 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.

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.[36] 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 what is socially acceptable behavior (norms),[37] 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).

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 indi-

[36] 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.

[37] 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.

viduals 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 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. (2020). Bordalo et al. (2016) argue that stereotypes are typically based on a "kernel of truth".[38] If one can demonstrate in a particular context that there are no statistical differences between two groups, this may induce a re-evaluation of the stereotype. 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 which 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 vs. no certificate).

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 differ-

[38]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.

ent contexts. The paper also demonstrates a central role for 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.

Bénabou, R. and J. Tirole (2011). Identity, morals, and taboos: Beliefs as assets. *Quarterly Journal of Economics 126*(2), 805–855.

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 (2020). Inaccurate statistical discrimination. *Mimeo*.

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.

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 (2020). The role of beliefs in driving gender discrimination. *Management Science (forthcoming)*.

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 (2022). Implicit preferences inferred from choice. Mimeo.

Danilov, A. and S. Saccardo (2017). Discrimination in disguise. Mimeo.

Darley, J. M. and P. H. Gross (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology 44*(1), 20.

Dovidio, J. F. and S. L. Gaertner (2004). Aversive racism. In M. P. Zanna (Ed.), *Advances in experimental social psychology*, Volume 36, pp. 1–52. Elsevier Academic 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.

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*.

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.

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. and M. R. Banaji (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological Review 102*(1), 4.

Greenwald, A. G. and L. H. Krieger (2006). Implicit bias: Scientific foundations. *California Law Review 94*(4), 945–967.

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. (2001). Description and prescription: How gender stereotypes prevent women's ascent up the organizational ladder. *Journal of Social Issues 57*, 657–674.

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.

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.

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 (2022). Non-bayesian statistical discrimination. *Mimeo*.

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.

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.

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.

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

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.

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.

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.

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

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.

A The Job Candidate Assessment: Tasks and Procedure

A.1 The word task (word certificate)

Participants solved three word search puzzles.[39] 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.[40] 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

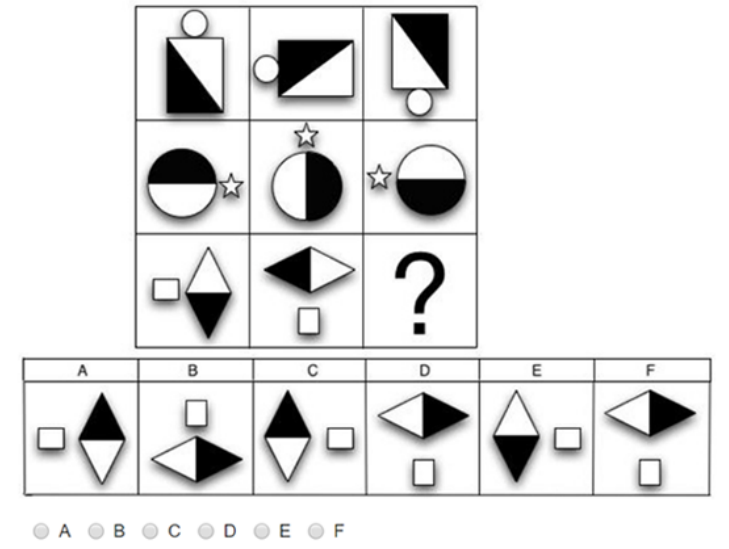
The order of the tasks was held constant for all participants. After the completion of all tasks, an incentivized belief elicitation and questions on demographics followed. For each of the tasks mentioned above, participants were asked how often they believed a randomly drawn

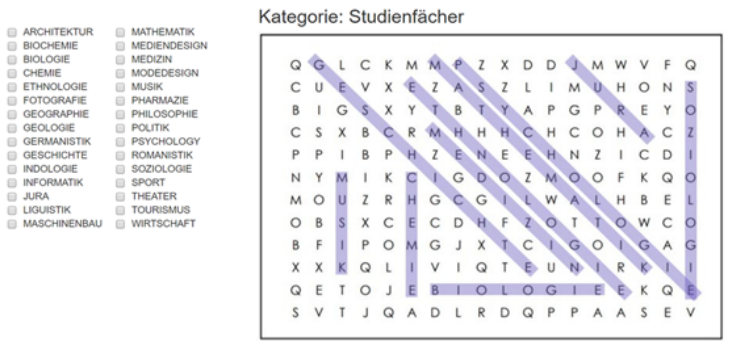
[39] Each puzzle had a theme: animals, countries or fruit.

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

male would perform better than a randomly drawn female.[41] They were also asked what 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 8: 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.





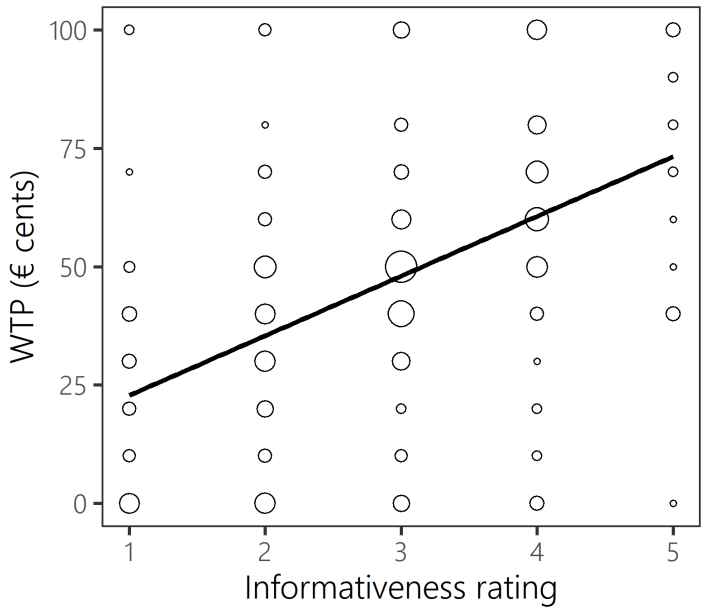
[41] More specifically, we asked participants to think about taking 100 draws of a pair of participants, each containing a randomly drawn male and a randomly drawn female from their session. They were asked to indicate how often they believed that the randomly drawn male performed better than the randomly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

B Additional Tables and Figures

Table 5: Demographic information of participants in the HIRING EXPERIMENT.

Variable	Treatment 1	Treatment 2
Age (mean, SD)	24.8 (5.4)	24.2 (4.4)
Gender: female (N, %)	59 (49.6)	60 (49.6)
Study subject: STEM (N, %)	56 (47.1)	65 (53.7)
Study subject: Economics/Business (N, %)	38 (31.9)	38 (31.4)

Figure 9: Correlation between methods of belief elicitation.



Notes: Scatter 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. Point size is proportional to number of responses. Line shows OLS fit.

Table 6: Aggregate results from all hiring decisions by treatment.

			Treatment 1			Treatment 2		
Decision	A	B	Pr. hire A init.	Pr. sell	Pr. keep A	Pr. hire A init.	Pr. sell	Pr. keep A
D_1 (T1)	(f,W)	(m,K)	0.513 (61/119)	0.176 (21/119)	0.395 (47/119)	–	–	–
D_1 (T2)	(f,K)	(m,W)	–	–	–	0.364 (44/121)	0.149 (18/121)	0.281 (34/121)
D_2	(f,K)	(m,K)	0.454 (54/119)	0.429 (51/119)	0.210 (25/119)	0.413 (50/121)	0.388 (47/121)	0.256 (31/121)
D_3	(f,W)	(m,W)	0.462 (55/119)	0.479 (57/119)	0.193 (23/119)	0.372 (45/121)	0.413 (50/121)	0.215 (26/121)
D_4	(f,W)	(f,K)	0.630 (75/119)	0.176 (21/119)	0.496 (59/119)	0.595 (72/121)	0.157 (19/121)	0.512 (62/121)
D_5	(m,W)	(m,K)	0.580 (69/119)	0.160 (19/119)	0.496 (59/119)	0.612 (74/121)	0.190 (23/121)	0.496 (60/121)
D_6	(f,K)	$(m,-)$	0.824 (98/119)	0.319 (38/119)	0.580 (69/119)	0.818 (99/121)	0.306 (37/121)	0.612 (74/121)
D_7	(f,W)	$(m,-)$	0.857 (102/119)	0.269 (32/119)	0.681 (81/119)	0.868 (105/121)	0.182 (22/121)	0.736 (89/121)
D_8	$(f,-)$	(m,K)	0.143 (17/119)	0.336 (40/119)	0.067 (8/119)	0.182 (22/121)	0.231 (28/121)	0.107 (13/121)
D_9	$(f,-)$	(m,W)	0.084 (10/119)	0.227 (27/119)	0.017 (2/119)	0.165 (20/121)	0.198 (24/121)	0.066 (8/121)

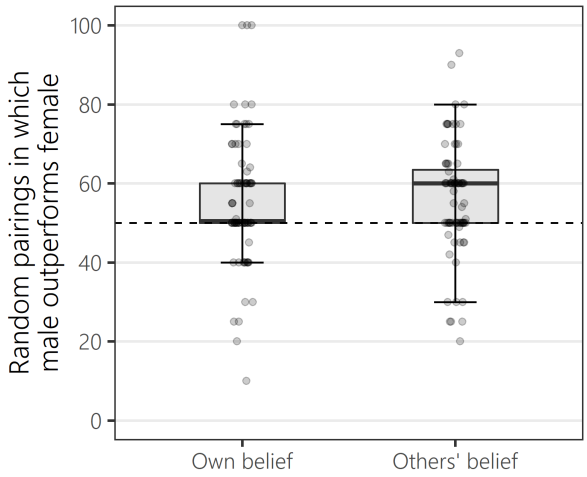
Notes: "Pr. hire A init." refers to hiring candidate A initially, "Pr. sell" refers to selling the initial choice (regardless of whether it was A or B), and "Pr. keep A" refers to hiring candidate A initially and not selling this initial choice. Hence, the difference between "Pr. hire A init." and "Pr. 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 D_1 (T1), 11.8% initially hire the woman and then sell the choice and 5.8% initially hire the man and then sell this choice; this means that conditional on hiring the woman the selling rate equals 23%, whereas it equals 11.9% conditional on hiring the man.

Table 7: Treatment effects on willingness to pay for qualifications.

	Full sample		By discrimination types (Multiple switchers imputed)			
	Multiple switchers excluded	Multiple switchers imputed	Explicit: against men	Explicit: against women	Explicit: mixed	Explicit: non
	(1a)	(1b)	(2a)	(2b)	(2c)	(2d)
T2	−2.4	−3.7	−21.0**	−3.2	−4.2	2.0
	(3.8)	(3.6)	(10.0)	(7.1)	(6.2)	(6.4)
Word	1.7	0.7	0.0	−3.2	−7.4	11.3*
	(3.9)	(3.6)	(12.5)	(8.3)	(5.0)	(5.9)
T2 × Word	7.3	8.1*	11.3	17.1*	6.4	3.8
	(4.9)	(4.6)	(14.1)	(9.4)	(7.5)	(7.8)
Constant	45.4***	46.9***	55.0***	45.5***	53.2***	38.9***
	(3.0)	(2.8)	(7.4)	(5.9)	(4.7)	(4.8)
N	438	480	62	124	152	142
R^2	0.015	0.014	0.095	0.044	0.010	0.071

Notes: (i) "T2" refers to Treatment 2 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. * p < 0.1, ** p < 0.05, *** p < 0.01.

Figure 10: 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.

Table 8: Descriptive statistics on actual performance in the job task 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

C 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?

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?

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.
