CHARLES UNIVERSITY

FACULTY OF SOCIAL SCIENCES

Center for Economic Research and Graduate Education

Dissertation Thesis

2024 **Daviti Jibuti**

CHARLES UNIVERSITY

FACULTY OF SOCIAL SCIENCES

Center for Economic Research and Graduate Education

Daviti Jibuti

Essays in Experimental Economics: Labor Market Discrimination

Dissertation Thesis

Prague 2024

Author: **Daviti Jibuti**

Supervisor: **prof. Randall K. Filer, Ph.D.**

Year of the defense: 2024

Dissertation Committee:

Prof. Randall K. Filer, Ph.D. (CERGE-EI, Charles University; Hunter College, City University of New York; Chair)

Andreas Menzel, Ph.D. (CERGE-EI, Charles University; Local Chair)

Doc. PhDr. Michal Bauer, Ph.D. (CERGE-EI, Charles University)

Referees:

Referee 1: Partha Deb, Ph.D. (Hunter College and the Graduate Center, City University of New York)

Referee 2: Daniel Hamermesh, Ph.D. (University of Texas at Austin)

References

Jibuti, Daviti. *Essays in Experimental Economics: Labor Market Discrimination.* Praha, 2023. 116 pages. Dissertation thesis (PhD.). Charles University, Faculty of Social Sciences, Center for Economic Research and Graduate Education – Economics Institute. Supervisor Prof. Randall K. Filer, Ph.D.

Abstract

The three chapters in this dissertation use field experiments to examine discrimination in various settings. Chapters I and II use a correspondence testing approach to study discrimination against applicants with visible tattoos in the German labor market. Previous empirical research has documented unfavorable treatment of tattooed applicants on the labor market. This may be because negative stereotypes are still associated with tattooed people, despite the increasing popularity of tattoos. However, the impact of tattoos on employment chances may be different across different occupations. Fictitious applications are sent to online job postings in the banking and IT sector. Otherwise identical applications differ only in the picture attached: in the treatment group the applicants have a visible tattoo. The extent of discrimination is measured by the difference in callback rates. The data indicates that the candidates without visible tattoos have, on average, a 13-percentage point higher callback rate in the banking sector, and in the IT sector applicants with visible tattoos are about 25% less likely to get a callback.

In the third chapter we conduct a study of hiring bias on an online platform where we ask participants to make hiring decisions for a mathematically intensive task. Our findings suggest hiring biases against Black workers and less attractive workers, and preferences towards Asian workers, female workers and more attractive workers. We also show that providing a candidate's information at the individual level and reducing the number of choices can reduce discrimination. On the other hand, provision of a candidate's information at the subgroup level further increases discrimination. The results have practical implications for designing better online freelance marketplaces.

Abstrakt

Ve třech kapitolách této disertační práce jsou použity terénní experimenty ke zkoumání diskriminace v různých prostředích. Kapitoly I a II využívají metodu korespondenčního testování ke studiu diskriminace uchazečů s viditelným tetováním na německém trhu práce. Předchozí empirický výzkum doložil nepříznivé zacházení s potetovanými uchazeči na trhu práce. Důvodem mohou být negativní stereotypy, které jsou s tetovanými lidmi stále spojovány, a to navzdory rostoucí popularitě tetování. Vliv tetování na šance na zaměstnání se však může u různých profesí lišit. Fiktivní žádosti jsou zasílány na internetové nabídky práce v bankovním a IT sektoru. Jinak totožné žádosti se liší pouze přiloženým obrázkem: v „treatment“ skupině mají uchazeči viditelné tetování. Rozsah diskriminace se měří rozdílem v míře zpětného volání. Z údajů vyplývá, že uchazeči bez viditelného tetování mají v bankovním sektoru v průměru o 13 procentních bodů vyšší míru zpětného volání a v IT sektoru mají uchazeči s viditelným tetováním přibližně o 25 % nižší pravděpodobnost, že dostanou zpětné volání.

Ve třetí kapitole provádíme studii předpojatosti při najímání zaměstnanců na online platformě, kde žádáme účastníky, aby se rozhodovali o najímání zaměstnanců pro matematicky náročnou úlohu. Naše zjištění naznačují předsudky při přijímání pracovníků proti černochům a méně atraktivním pracovníkům a preference vůči asijským pracovníkům, ženám a atraktivnějším pracovníkům. Ukazujeme také, že poskytování informací o uchazeči na individuální úrovni a snížení počtu možností volby může diskriminaci snížit. Na druhou stranu poskytování informací o uchazeči na úrovni podskupin diskriminaci dále zvyšuje. Výsledky mají praktické důsledky pro navrhování lepších online freelance platforem.

Keywords

Labor market discrimination; field experiment; visible tattoo; gig economy; gender discrimination

Klíčová slova

Diskriminace na trhu práce; terénní experiment; viditelné tetování; gig ekonomika; diskriminace na základě pohlaví

Length of the work: 163, 858 characters with spaces, without abstract and appendices

Declaration

1. I hereby declare that I have compiled this thesis using the listed literature and resources only.
2. I hereby declare that my thesis has not been used to gain any other academic title.
3. I fully agree to my work being used for study and scientific purposes.

In Prague on Daviti Jibuti

18.11.2023

Acknowledgement

I am grateful to many people who have provided help and support throughout my dissertation work, and I want to express my gratitude to all of them.

First of all, I would like to express my great appreciation to my supervisors, Randall Filer and Andreas Menzel for their insightful comments and suggestions and continued encouragement. Both believed in me and my research topic and provided patient guidance to help me successfully finish my studies.

Second, I am particularly grateful to Michal Bauer, a member of my dissertation committee. He has devoted lots of his valuable time to review my papers and his suggestions helped to improve the quality of my work and broaden my knowledge in the area of experimental economics.

Part of this dissertation was written during my research stay at the University of Chicago in the Fall of 2017. John List was extremely generous to invite me to Chicago and spend few months with his research team. Discussions with John and presenting my work in a research seminar gave me many useful insights and I got very valuable feedback. Moreover, I had an opportunity to discuss my work with faculty members at the University of Chicago and Booth School of Business.

This dissertation would not be complete without my co-authors, who I met at the University of Chicago. As both of us were working in the same field, Weiwen Leung offered me the opportunity to co-author the paper that comprises the third chapter of this dissertation. I would like to thank Weiwen for his generosity and helpful suggestions on my work.

I have also benefited from a number of discussions with many other excellent academics. I would like to thank Patrick Gaule, Fabio Michelucci, Peter Katuscak, Nikolas Mittag, Gerard Roland, Stepan Jurajda, Fatemeh Momeni, Eszter Czibor, Jan Fidrmuc, Mariola Pytlykova, Jan Kmenta, Alex Imas and participants of various conferences, seminars, and workshops.

I extend my thanks to my friends and colleagues Nikoloz Kudashvili, Gega Todua, Ala Avoyan, Danijela Vuletic, Dejan Kovac and Patrick Nüß for their encouragement, helpful discussions and keeping me motivated.

I would also like to express my gratitude to the CERGE-EI Academic Skills Center, who polished my writing. Namely, I would like to thank Andrea Downing, Deborah Novakova and Paul Whitaker for their hard work and patience towards me.

I would especially like to thank my parents and my brother for their unconditional love, support, encouragement and patience. I must also thank my wonderful wife, Monika, who has been extremely supportive throughout this entire journey and made countless sacrifices for me.

All errors remaining in this text are the responsibility of the author.

Prague, Czech Republic Daviti Jibuti

November 2023

Table of Contents

Introduction

In my dissertation, I study labor market discrimination and run experiments to collect and analyze data. The first two chapters focus on discrimination against applicants with visible tattoos in different settings and examine the sources of such discrimination. The third chapter studies how race, gender and beauty of the potential applicants, as well as the number of available applicants, affect the hiring decisions of employers. I describe each chapter in more detail below and explain their contribution to the literature.

The first two chapters, which contribute to the experimental discrimination literature, study labor market discrimination against applicants with a visible (neck) tattoo. Despite tattoos becoming increasingly common in every group of society (Swami et al., 2015), negative stereotypes are still associated with tattooed individuals (Timming, 2015). This may translate into an unfavorable treatment of applicants with visible tattoos in the labor market, as HR managers may relate tattooed candidates to low productivity workers, hence statistically discriminating against them (Phelps, 1972). Besides statistical discrimination, tattooed applicants may also face taste-based discrimination (Becker 1957), a form of discrimination driven by personal preferences (in this case, distaste of tattoos). So far, there is little known about how visible tattoos can affect employment chances, and the first two chapters of my dissertation provide experimental evidence of such discrimination, therefore contributing to the small, but recently growing stream of literature which examines objects of discrimination that are not exogenously given to the candidate.1

I use the correspondence testing method (Jowell and Prescott-Clarke, 1970) to collect data and conduct two natural field experiments in Germany, where including a photograph in the job application is a common practice. The first chapter provides evidence of discrimination against tattooed applicants in the banking sector: having a visible tattoo reduces callback probability by about 64% and the effect remains largely stable even after controlling for numerous covariates. I attempt to examine whether tattooed applicants face statistical discrimination (Phelps, 1972) by providing additional information in the application. Additionally, the design of my experiment allows testing for customer/coworker discrimination. However, I was unable to find supportive evidence

[1] Exogenously given characteristics include age, gender, race/ethnicity, etc., while other characteristics, that are the choice of the individual include religion, military service affiliation, marital status, motherhood and job changes, among others.

in my data for these theories and argue that tattooed applicants may face taste-based discrimination (Becker, 1975) in the banking sector.

The second chapter examines whether discrimination found in the banking sector is a sector-specific phenomenon or applies also to some completely different settings. Accordingly, I examine the IT sector, which is a more skill-intensive sector in which workers usually have minimal face-to-face interaction with customers. Additionally, to address Heckman and Siegel's (1993) critique of correspondence studies, I randomly vary the skill level of my fictitious applicants, as suggested by Neumark (2012). This also helps me to test whether statistical discrimination could be the reason behind the unfavorable treatment of tattooed applicants. I find that applicants with visible tattoos have a 25% less chance of getting a callback, however, as in the previous chapter I did not find evidence of statistical discrimination. Overall, by combining the results of these chapters I argue that visible tattoos seem to be broadly disliked in the labor market in Germany.

In addition to expanding the literature by examining discrimination against choice-based characteristics, studying discrimination against applicants with visible tattoos may have additional policy implications. A majority of correspondence studies in the discrimination literature focuses on characteristics that are exogenous to the applicant (Bertrand and Duflo, 2017), as discrimination based on these characteristics is usually prohibited by law. However, labor markets are becoming increasingly diverse, particularly in developed countries (OECD, 2020) which prompts policymakers to extend laws and protect various groups against discrimination. For example, in the US, among other traits such as race, gender, age, etc., unequal treatment is prohibited against political or union affiliation and physical appearance (Baert 2018). In this context, studying discrimination against choice-based characteristics further expands policymakers' understanding of the extent of discrimination against various groups, and can help them to prevent it.

The third chapter, co-authored with Weiwen Leung, Zheng Zhang, Jinhao Zhao, Maximillian Klein, Casey Pierce, Lionel Robert and Haiyi Zhu, contributes to several streams of literature. By studying how hiring decisions are affected by the race, gender and beauty of potential candidates, we contribute to the discrimination literature. We also examine how the number of available candidates that apply for a job affects hiring decisions, which is related to the choice overload literature. We conducted an online experiment on Amazon MTurk to examine discrimination across different dimensions.

Besides uncovering some racial discrimination (Asians preferred over Whites and Whites preferred over Blacks), our data also suggest the existence of a beauty premium. In terms of sources of discrimination, we find evidence of statistical discrimination in favor of certain groups (Asians, and physically more attractive candidates). Moreover, we show that providing certain types of additional information and reducing the number of available candidates from whom the employer must chose can reduce hiring bias. We argue that our findings may have practical implications in reducing hiring bias in an online marketplace.

1. Discrimination against Workers with Visible Tattoos: Experimental Evidence from Germany[2]

1.1 Introduction

Correspondence studies have often been used to detect discrimination in labor markets (Neumark, 2016). Most of these studies have focused on characteristics that are exogenously given to the person, such as race, gender, ethnicity, age and etc. (see Bertrand and Duflo (2017) for the most recent review of field experiments studying discrimination in various settings). Some recent papers have examined objects of discrimination that some view as an exogenously given characteristic, but which others see as the choice of the individual; for example, sexual orientation (Gneezy, List and Price, 2012) and wearing a hijab (Weichselbaumer, 2019). So far, little is known about the impact of purely choice based characteristics on employment opportunities. In this paper I look at how visible tattoos, which are an individual choice (French, Mortensen and Timming, 2019), affect employment opportunities.

Studying discrimination against applicants with visible tattoos provides a novel way to measure the costs people pay to express their identity. Based on the Social Identity Theory (Tajfel and Turner, 1979) Akerlof and Kranton (2000) introduced identity into economics. They argue that the concept of identity expands analyses in multiple domains and claim that "identity can explain behavior that appears detrimental" (Akerlof and Kranton(2000) p. 717). Unlike the "standard economic agent", incorporating identity (or self-image) into the utility function changes the economic analyses and explains many unexplained facts. Namely, identity can help to explain why people behave in ways that may seem maladaptive to others. In the Akerlof and Kranton (2000) model, deviation from prescriptions for behavior that "mimics the ideal" from your social category may lead to "punishment" from society and disutility from this deviation. The authors cite

---

[2] This work was published in Jibuti (2018) „Discrimination against Workers with Visible Tattoos: Experimental Evidence from Germany", CERGE-EI Working Paper Series No. 628. I am grateful to Randall Filer and Andreas Menzel for their continuous support and guidance. I would like to thank Michal Bauer, Patrick Gaule, Fabio Michelucci, Peter Katuscak, Nikolas Mittag, Gerard Roland and conference/seminar participants at the University of Chicago, University of East Anglia, the University of Rennes 1 and Masaryk University for their valuable comments and helpful suggestions. Bilal Zafar, Christian Scherer and the whole team of richtiggutbewerben.de provided qualitative applications that guaranteed a high response rate. All remaining errors are mine.

tattooing and body piercing as extreme examples of expressions of identity. While this behavior may match an ideal within the group, according to the model it has an adverse effect outside of the group. As a result, they might be treated less favorably when interacting with individuals from different groups.[3] One of the most important interactions occurs in the labor market, such as searching for a job, or communicating with coworkers or customers. Therefore, Akerlof and Kranton's (2000) model predicts that individuals with tattoos may find it difficult to succeed in labor markets where most people do not have tattoos and do not endorse them.

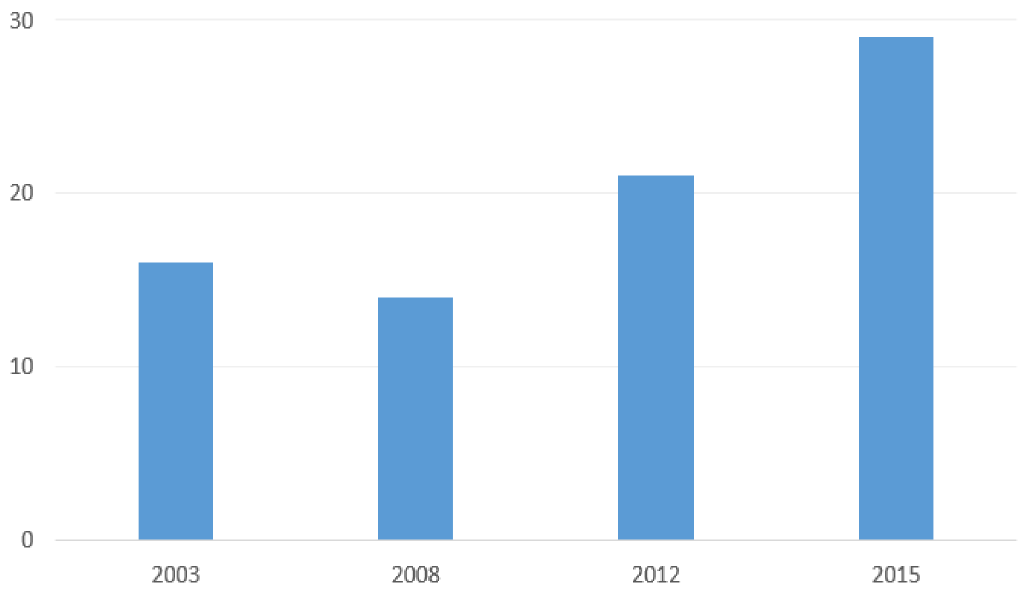
While in many instances the choice of belonging to a certain social group may not be observable, some choose to visibly express their group identity. Examples include wearing a hijab (Weichselbaumer, 2019) and tattoos (Akerlof and Kranton, 2000). There may be two types of people: those who are unaware that visible tattoos may impact their employment opportunities, and those who are aware but prioritize group identity. However, their choices may come at a cost. HR managers might see visible signs of membership in a particular social group as a hindrance to employment, due to their own or society's distaste for a particular group. They may also think that an individual belonging to a particular social group may be less productive than others. Thus, as predicted by Akerlof and Kranton's (2000) model as well as Social Identity Theory, HR managers may treat those individuals less favorably who explicitly express their social group identity. So far, there is very little rigorous research on the impact of tattoos (as one of the forms to express identity) on labor market outcomes. This paper aims to shed light on the issue by running a natural field experiment to test how having a visible tattoo on an application photo affects employment opportunities.

The research question is also relevant given that tattoos are becoming increasingly popular. For example, Figure 1.1 shows that almost 30% of the US population in 2015 had at least one tattoo, of which 30% are visible.[4] Aslam and Owen (2013) argue that almost 25 % of the adult population in Europe today have tattoos. Akerlof and Kranton (2000) note that "tattooing, body-piercing, ... all yield physical markers of belonging to more or less explicit social categories and groups" (p. 721). These groups would be different than they used to be, for example 30, or even 20 years ago. Historically, tattoos

[3] Results of Akerlof and Kranton (2000) model are consistent with Social Identity Theory, which states that affiliating with a social group result in in-group favoritism and out-group discrimination (Tajfel and Turner, 1979).
[4] Source: The Harris Poll #12, February 2016.

were associated with sailors, prisoners and gang members (Timming, 2015), who would possibly be less productive in the workplace, especially for white-collar jobs. However, today tattoos are becoming mainstream and are more accepted by wider social groups (Antonellis et al., 2017). Despite this, there is mixed evidence on the correlation between tattoo status and personal characteristics that are essential for social interactions or may be correlated with productivity. Ruffle and Wilson (2018) conducted an experiment on Amazon Mechanical Turk (or Mturk) and found that tattooed and non-tattooed individuals were equally honest. On the other hand, using the Mturk survey of 2008 subjects, Mortensen, French and Timming (2019) report that tattoos are negatively correlated with health-related outcomes (mental health and trouble sleeping) and risky behaviors (smoking, being in jail or prison and number of sex partners past year). All these characteristics are arguably correlated with productivity. Therefore, it seems plausible that many people consider tattoos as a signal of individual characteristics that are relevant in the labor market, resulting in an unfavorable treatment of tattooed applicants.

Figure 1.1: Tattooed individuals in the US (all age groups), %



In addition to signaling group identity, visible tattoos may also have an effect on applicants' employment chances through their appearance. It is well documented that noncognitive personal attributes affect labor market outcomes, including hiring, firing and promotion. A number of studies have shown negative bias towards less physically attractive applicants (Hamermesh and Biddle, 1994; Harper, 2000; Weichselbaumer, 2019; Dillingh, Kooreman and Potters, 2016; Katuscak and Kraft, 2013; Ruffle and Shtudiner, 2015). The relationship between appearance and labor market outcomes might exist through the preferences (of HR managers) or an expected correlation between appearance and productivity. Attractiveness can refer to given features of an individual that are hard to change. But other aspects of one's appearance, such as clothing style, can be changed at almost zero cost. Having a visible tattoo is not costless, however, as once acquired, it is hard to change.

Despite the growing popularity of tattoos (Stirn et al., 2006), prejudices related to them still exist in society. As Timming (2015) points out, tattoos are linked to negative stereotypes, including promiscuity, crime, drug usage, decreased honesty, low levels of generosity and intelligence, and gang membership. While some stereotypes might indeed be accurate to some degree ("the Dutch are tall"), many are mostly false ("Florida residents are elderly") (Bordalo, Coffman, Gennaioli and Shleifer, 2016). Bordalo et al. (2016) build a model of stereotypes in which decision makers overweight a group's most distinctive types when making predictions about that group. They show that stereotyping increases the systematic differences made between groups, even when the actual difference is negligible. In our context, when visibly tattooed individuals apply for a job, hiring managers may relate them to groups with low expected productivity, resulting in lower hiring rates of tattooed applicants. In the economics literature, this form of discrimination is referred to as statistical discrimination (Phelps, 1972), as decision-makers try to navigate through imperfect information. Alternatively, employers may have a simple distaste of visibly tattooed workers, leading tattooed individuals to have lower hiring rates simply due to taste-based discrimination (Becker, 1957).[5]

Based on the above considerations, I expect a visible tattoo to have an adverse effect on employment chances: job applications submitted with a visible tattoo on the photo would generate significantly less call-back rate compared to the same application but the photo without a visible tattoo. Additionally, empirical evidence suggests that visible tattoos are a hindrance factor in customer-facing positions (Baumann, Timming and Gollan, 2015), hence I conjecture that discrimination will increase in front-office

[5] Having a visible tattoo will potentially have a heterogeneous effect across occupations and demographic groups. In marketing, for example, it might bolster an individual's opportunities as it can be seen as a signal of creativity. On the other hand, in a more "conservative" occupation, such as banking (which is our focus in this paper), it may hinder employment opportunities, as hiring managers might be worried that visible tattoos are not accepted in this occupation because of consumers' and/or coworkers' perceptions.

positions and positions requiring teamwork as in addition to their own preferences, HR managers would have a concern that customers or co-workers would not like to interact with visibly tattooed employers. Furthermore, providing more detailed information regarding the applicant should help HR managers adjust their inaccurate beliefs about the expected productivity of tattooed applicants, and hence reduce/eliminate statistical discrimination.

I do not claim to study the effects of all types of visible tattoos on employment opportunities in a broad range of sectors. Specifically, I examine how a "neutral" visible tattoo (a tribal tattoo[6]) affects employment opportunities in the banking industry. Although I find evidence of differential treatment, this does not imply that any type of visible tattoo will have the same effect across industries. Tattoos that elicit positive emotions may even help applicants in some fields such as marketing, sports, or fashion. For example, using the Amazon MTurk study Timming (2017) found that while a (neutral) visible tattoo adversely affected employment chances in the fine dining industry, it was a significant asset for the positions at a nightclub. Therefore, considering a broader range of occupations and the type of tattoos will help to generate generalizable results. Our paper attempts to answer two main questions. First, do individuals with visible tattoos face discrimination in the banking labor market? Second, what are the sources of discrimination? To answer these questions, I run a natural field experiment in Germany. I design fictitious applications and send them to online job advertisements in the banking sector. Applications include a picture of the candidate, and the tattoo is present in the picture in the treatment group. Callback rates are recorded and analyzed as to whether they differ for candidates with and without visible tattoos. In response to the research questions, I find the following. First, applicants without visible tattoos have on average 54% higher callback rates (with $p - value = 0.0001$). Second, using two channels that provide a positive signal about the personality of the applicant, I fail to find evidence of statistical discrimination. In addition to a lower callback rate, discrimination may have a different form. In particular, I find that employers respond positively to candidates without tattoos significantly faster than to those with tattoos. Our data do not suggest that visibly tattooed applicants face coworker or customer taste-based discrimination. I control for factors such as the degree of interaction with coworkers and customers. These factors

[6] Timming and Perrett (2017) classify tribal tattoos as “neutral in content”, hence we use tribal tattoos in our study.

are used to test the model of (coworker and customer) taste-based discrimination; however, none of the interactions with a treatment dummy are significant in regression analyses.

Only a limited number of studies have reported a correlation between visible tattoos and negative labor market outcomes, particularly in relation to employment. By analyzing panel data on Dutch individuals, Dillingh et al. (2016) argue that visibly tattooed candidates score less favorably, in particular on health items (physical as well as mental), though the relationship in the case of the labor market is relatively weak. Similar conclusions are drawn in work by French, Maclean, Robins, Sayed and Shiferaw (2016), who use two large data sets from the US and Australia. They report that after controlling for personal characteristics, candidates with visible tattoos are treated similarly in the labor market to candidates without them. Overall, both studies mentioned above fail to establish a correlation between tattoos and the labor market outcomes (employment and wages).[7] On the other hand, there is some evidence of a negative relation between tattoos and employment in the sociology literature. Relying on 25 in-depth interviews with managers and tattooed workers, Timming (2015) concludes that there is a negative bias toward candidates with visible tattoos. By conducting online experiment Timming et al. (2017) argue that visible body art may hinder employment opportunities. The authors choose 120 respondents (from 182 overall) to show them photographs and asked how likely they would hire the person depicted in the picture. Photographs were experimentally manipulated to show a person without tattoo/piercing (control group), with tattoo and with piercing. The authors show that tattoo and piercing reduced the likelihood that person would be hired, though this effect was lower for the non-customer-facing roles (Timming et al., 2017). Brallier, Maguire, Smith and Palm (2011) focus on the restaurant industry and show that 88% of managers are willing to hire applicants without visible tattoos, while only 70% of managers are willing to hire applicants with tattoos. Miller, Nicols and Eure (2009) show that workers without body art prefer not to work alongside colleagues with visible body art (tattoo(s) and piercing(s)). Swanger (2006) reports that around 90% of hiring managers surveyed in the hospitality industry

[7] I should note that these studies consider a broad range of occupations. Finding no correlation might indicate that visible tattoos may not have a homogeneous effect across different occupations.

claim that individuals with tattoos and/or piercings are viewed negatively by managers. Consequently, they are less willing to hire these individuals.[8]

These studies of the sociology literature typically use laboratory experiments, conducted in a tightly controlled environment, which raises concerns about the experimenter demand effect and its influence on results and external validity of findings (List, 2006). List (2008) argues that natural field experiments are a useful tool to address this concern, as subjects in the natural field experiments are not aware of their participation in the experiment, hence their behavior would not differ from natural behavior. Consequently, a well-designed field experiment provides a setting most closely replicating the actual behavior, hence potentially the best way to test economic theories (Levitt and List, 2007). Therefore, building on those studies mentioned in the previous paragraph, I test the presence of discrimination against tattooed applicants among actual employers making hiring decisions without the knowledge of being a part of the experiment. I am not aware of any natural field experiment in economics literature examining the discrimination of tattooed applicants. The paucity of evidence about hiring bias for tattooed workers limits our knowledge about the extent (and the source) of discrimination against them.

Most of the experimental correspondence studies in the field focus on characteristics of the candidate that are given, for example, race/ethnicity (Bertrand and Mullainathan, 2004; Bartos, Bauer, Chytilova and Matejka (2016); Kaas and Manger, 2012), gender (Riach and Rich, 1987, 2006) or disability (Riach and Rich, 2002, Gneezy, List and Price, 2012). However, literature relating controllable characteristics to discrimination is scarce. Our study is most closely related to work by Gneezy et al. (2012), Weichselbaumer (2019) and Cohn, Marechal, Schneider and Weber (2017). Among other characteristics, Gneezy et al. (2012) examine discrimination against sexual minorities in the car repair market, while Weichselbaumer (2019) tests labor market discrimination against female ethnic minorities (Muslims) wearing a headscarf in Germany. As Gneezy et al. (2012) note, some individuals perceive sexual orientation as a personal choice, while others think that it is a given characteristic (such as race or age). The authors claim that if the decision maker believes the object of discrimination is a given characteristic,

[8] Ozane et al. (2019) also conducted an experiment on Amazon MTurk to examine the impact of tattoos in the service failure context. Contrary to the studies mentioned above, the authors find that tattoos did not have an influence on customers' propensity to generate negative word-of-mouth.

discrimination is taste based. Using several treatment manipulations, Gneezy et al. (2012) find that ethnic minority car dealers make significantly higher price offers to a gay couple while Caucasian dealers treat them favorably. Similarly, Weichselbaumer (2019) argues that it is not clear (at least to HR managers in Germany), whether the hijab is a choice made by Muslim women or if they are forced to wear them. The author runs a natural field experiment in Germany and finds that female applicants with a Turkish name and hijab face a high level of discrimination that is mostly motivated by animus.

While there is some ambiguity in the perception of sexual minorities or women wearing a hijab, frequent job switching is an individual choice. Cohn et al. (2017) examine how frequent job changes affect employment chances. They conduct laboratory, field and survey experiments and find that workers who change jobs frequently have a lower chance of employment compared to more "stable" workers. The possible reason is that more stable workers have (or at least are perceived to have) better non-cognitive skills and therefore more firms are willing to hire them. I should note that if having a visible tattoo is not necessarily correlated to a person being less productive, frequent job changes may have a direct effect on productivity. For example, switching jobs regularly between industries might affect the productivity of the worker negatively through less accumulation of industry-specific human capital. Therefore, in the eyes of employers, frequent job changes, particularly between industries, might signal lower productivity, and hence a lower chance of employment. This is confirmed in the study by Cohn et al. (2017). Using a field experiment, they find that having relevant industry experience increases the chance of a callback and almost offsets the negative impact of frequent job changes.

Despite some exceptions as discussed above, there is still limited research relating choice-based characteristics to labor market outcomes. Tattoos are outcome of individual choice (French et al., 2019), particularly in the setting of the current study.[9] This also distinguishes our paper from other correspondence studies in the field that examine discrimination against exogenously given characteristics. To the best of our knowledge, this is the first field experiment that examines discrimination against applicants with visible tattoos in the banking industry, thus complementing the existing literature.

[9] I should note, however, that I cannot rule out that due to cultural reasons this choice may be sometimes affected by social pressure or parental choices (for example in indigenous populations).

Studying discrimination against tattooed individuals in the labor market could be useful for policymakers as well as for the public. Currently, it is up to a company to ban visible tattoos in the workplace. However, policymakers could publicize (at least, among employers) that tattoos are becoming mainstream, and thus lessen negative stereotypes associated with tattooed workers. This could potentially reduce discrimination. On the other hand, individuals who plan to work in a sector where having a visible tattoo might diminish their employment chances, might reconsider their decision to have a tattoo on a visible part of the body. Alternatively, individuals might view visible tattoos as signaling that they are not willing to work in an environment where tattoos (or expressions of identity) are not accepted.

The rest of the paper is organized as follows. The next section presents the design of the experiment, including description of applications, the application process and summary statistics of the sample. The results are discussed in Section 3, while Section 4 concludes.

1.2 Experimental Design

I use a correspondence study to collect data, in line with the literature. The correspondence testing approach, in which fictitious CVs are sent to real job vacancies, was developed by Jowell and Prescott-Clarke (1970). This approach proved successful in examining discrimination in the labor market (Bertrand and Duflo, 2017). The candidates' CVs are identical, and they exist only on paper/electronically, except for the variable of interest, for example, ethnicity, gender or sexual orientation. Discrimination is measured by the interview invitation rate of the applicants. While the correspondence testing approach has many advantages over other practices, it also has limitations (Riach and Rich, 2002). An invitation for an interview is not the final outcome of the recruiting process; thus I cannot observe hiring rates or wages offered for the particular candidate. However, Riach and Rich (2002) analyze discrimination at the interview invitation and job offer stages and find that around 90 % of discrimination occurs at the stage of selecting candidates for an interview; thus, the callback is a key part of the process. Another drawback of the method is that one cannot use it in sectors in which written applications are not typically used (for example low skilled jobs). Despite these limitations, the correspondence testing approach has obvious advantages over other techniques and is widely used by researchers. One advantage is that at a much lower cost than for audit studies,[10] a researcher can obtain larger sample sizes. Another advantage is that correspondence studies avoid experimenter effects that may lead to biased results in audit studies.

The reason I chose Germany for the study is twofold. First, inclusion of a photo in job applications is common in Germany. Another reason is the relatively large size of the labor market. According to the Federal Statistical Office in August 2016, there were 685,238 vacancy announcements throughout the country.[11] For these reasons, the German labor market has been very popular among scholars examining discrimination against candidates with different characteristics. For example, researchers have been able to document discrimination against ethnic minorities (see, for example, Bartos et al., 2016; Weichselbaumer, 2019; Kaas et al., 2012; Goldberg, Mourinho and Kulke, 1996) and unattractive candidates (Katuscak et al., 2013).

1.2.1 Applications

Following the standard German labor market practice, each application included a cover letter and a CV with a photo of the candidate. In order to generate sufficient data[12] for the analysis, I needed to have sufficient callback rates. Therefore, I asked HR professionals to create applications in accordance with German standards.[13] All applicants in our study are the same age (born in 1989). I use German-sounding names and surnames for our applicants to rule out potential discrimination toward minorities, which is well documented in Germany (Kaas and Magner, 2012; Bartos et al., (2016)).

As Weichselbaumer (2019) claims, sending more than one application to an employer may bias results, as the method of testing for discrimination is increasingly well known among HR managers in Germany. To avoid this problem, I respond to each job advertisement with only one application. Thus, the name and contact details are the same

---

[10] In audit studies, two auditors/testers (often actors) are matched in terms of all relevant observable characteristics, including physical appearance, except for the object of discrimination. They apply for the job position and – differently from correspondence studies - if invited, physically go to the interview, while being instructed to "behave similarly" (i.e., act, talk, and dress similarly). This differs from correspondence studies in which recruiters only see the CVs of applicants, and when, if fictious applicants are invited for an interview, the invitation is declined (Neumark, 2012).
[11] The number is taken from http://www.statistik-portal.de/Statistik-Portal/en/en_zs02_bund.asp
[12] I conducted power calculations prior to the experiment in order to target sample size.
[13] Sample CVs are available from the author upon request.

for both tattooed and non-tattooed candidates.[14] As I send applications to different firms, I can use exactly the same application for both tattooed and non-tattooed candidates.[15] I use two channels to test the model of statistical discrimination. In the indirect channel, contact details of a non-existing referee are included in a fraction of applications. If the HR manager wanted to acquire more information regarding the candidate, she could contact the referee via email or phone. To minimize the cost to the employer, the job applicants and referee did not pick up the phone; however, I recorded the missed calls. If an employer contacted the reference person via email, I replied (on behalf of a referee) by describing the candidate as reliable, highly consumer oriented and a team player. HR managers might not invest time searching for additional information about the applicant, using just the information available from the application itself, and may not contact the referee at all. For this reason, I use another channel whereby I indicate on the applications that the applicant is a member of the local alpinist association (Deutsche Alpenverein) which is well known and popular in Germany. I conjecture that being a member of this group is associated with that person being trustworthy, reliable and a team player.[16] If the additional information provided by a referee or by group membership reduces discrimination, the results will be consistent with the model of statistical discrimination.[17] Further, by controlling the degree of interaction with consumers and/or colleagues, my experimental design enables me to test consumer and coworker taste-based discrimination (Becker, 1971).

Finally, I chose several candidates for the photo. I ran a survey of undergraduate and graduate students studying in various universities in Prague (Czech Republic) with an initial pool of ten pictures. Based on the survey results, the experiment subjects were ranked in the following attractiveness categories: below average looking, average looking and above average looking. From each category I choose one picture of male and female faces, ending up with a total of six pictures.[18] Pictures from different categories are randomly assigned to applications. In the treatment group, I added a tattoo to the picture

[14] Obviously, I use a different first name for male and female candidates.
[15] The only difference is in the photo of the applicant: in the treatment group, candidates have a visible tattoo.
[16] By reviewing the psychology literature on mountaineers, Jackman et al. (2020) stated that the assessment of agreeableness (one of the big five personality traits) of mountain climbers is inconclusive. On the other hand, the literature does find that having a trusted climbing partner is crucial for reducing risk.
[17] Here I only consider channels of statistical discrimination that relate to the personality of the candidate. There might be other sources of statistical discrimination that I do not consider in this paper.
[18] Appendix A for detailed discussion of the survey.

using computer software. This rules out possible concerns related to the quality of the photo. I chose different tattoos for male and female applicants. The photo of the candidate is on the cover page of the application, to ensure that our signal (having a tattoo) reached the employers. Within the gender and attractiveness sub-groups of applications, the only difference between the applications in treatment and control group is the photo of the candidate. Therefore, any difference in callback rates should be associated with the tattoo.[19]

1.2.2 The Application Process and Data Description

I focus on job openings across Germany in the banking sector for several reasons. First, there is a sufficient number of new openings in this sector. Second, there are both front- and back-office positions in banks, which enables us to test whether tattooed applicants face customer discrimination (perceived by employers). The banking sector also features variation in firm characteristics. For example, in terms of age, there are firms in our sample less than one-year-old along with very old firms aged 150 and even more. There is also variation in firm size, defined as the number of employees. Further, our sample includes both international and domestic firms.[20] Therefore, because I use between-firm design, I make sure that the firms are similar in terms of those characteristics for tattooed and non-tattooed applicants. To summarize, I focus on the roles with direct/indirect customer support for existing or new financial products and advisors in financial matters, including investment and financing.

Table 1.1 presents summary statistics of the firms' characteristics as well as job specific requirements.[21] The average age of the firms is 34 years (with s.d.=35.95), and average size is 1298 employers (with s.d.=4489.37). 37% (with s.d.=0.48) of the firms operate internationally while only 14% (with s.d.=0.35) are recruitment agencies recruiting workers on behalf of others. As our aim is to compare the treatment effect

[19] I should note that our results are context specific, conditional on "neutral tattoos", specifically, a tribal tattoo. If the content of the tattoo is specific, for example, army related, the results might be different

[20] The data on firms' characteristics (age, size, international status, etc.) was collected either directly from their web pages or through social network accounts, such as LinkedIn or Xing.

[21] The third and fourth column of Table 1 shows the means of the firm/job characteristics across the treatment and the control group and the last column shows a p-value of the hypothesis of equal means across different groups. None of the differences is significant; hence, I can claim that any differential treatment should be due to the treatment itself. Further, I ensure that regions were also balanced across groups. See Appendix B for a discussion of the randomization check for regions.

across the front and back-office positions, 43% (with s.d.=0.49) of our sample had advertisements for the front office. Most of the job postings (87% with s.d.=0.33) were collected in urban areas.[22] Each firm posted on average 17 (with s.d.=78.16) advertisements in a particular job portal. Most of the applications, 81% (with s.d.=0.39), were sent via email, while the rest were submitted through the online application system. As mentioned above, firms had a variety of requirements. In particular, teamwork was required by 46% (with s.d.=0.49) of firms[23], while a neat and friendly appearance was explicitly required in 25% (with s.d.=0.44) of cases. Almost half of the firms (48% (with s.d.=0.49)) required that the candidate should indicate an expected/required salary in the application.

Table 1.1: Summary statistics

Firm/job characteristic	*Mean*	*Obs.*	*Non-Tattooed*	*Tattooed*	*P-value*
Firm size	1298 (4489.37)	782	1268 (5218.41)	1327 (3652.49)	0.85
Firm age	34 (35.95)	782	34 (36.87)	34 (35.07)	0.76
International firm	0.37 (0.48)	782	0.36 (0.48)	0.38 (0.49)	0.58
Number of ads	17 (78.16)	782	16 (75.45)	18 (80.77)	0.62
Teamwork requirement	0.46 (0.49)	610	0.44 (0.49)	0.47 (0.50)	0.39
Neat/Friendly appearance requirement	0.25 (0.44)	782	0.25 (0.43)	0.26 (0.44)	0.75
Submit email vs. online	0.81 (0.39)	782	0.81 (0.40)	0.83 (0.39)	0.51
Front office	0.43 (0.49)	782	0.44 (0.49)	0.42 (0.49)	0.61
Urban area	0.87 (0.33)	782	0.86 (0.35)	0.88 (0.32)	0.31
Recruitment agency	0.14 (0.35)	782	0.15 (0.36)	0.14 (0.34)	0.56

Notes: The table shows means of firm/position characteristics and their comparison across the treatment (tattooed) and the control (non-tattooed) groups. Standard deviations are in parenthesis beneath mean estimates. The last column shows p-values of the hypothesis of equal means across groups.

The experiment was conducted in Germany from October 2016 to the end of January 2018.[24] Some keywords were used to search for relevant job advertisements in

[22] I define the area as urban if the population of that area is greater than 100 000.

[23] I started to collect the data for the dummy variable "teamwork" at a later stage of the experiment, resulting in a lower number of observations compared to other groups.

[24] With some interruptions the data collection process lasted 12 months.

the most common job portals in Germany.[25] I combined application materials in a single file and responded to the job postings. Some firms also required various certificates (from university and/or previous jobs). However, for objective reasons, it is difficult to create fake certificates. Thus, applications consisted of a CV with a picture and a cover letter. A number of employers requested certificates by contacting the candidate through email, in order to have a complete application. In those cases, I replied that for organizational reasons it was not possible to send certificates and the candidate would provide them during the interview. Overall, only 5% of employers required certificates and requests did not differ across the treatment and control group.

When employers requested applicants to indicate salary expectations, I identified the regional location of the workplace and indicated average earnings in the "financial and insurance activities" sector from that region.[26] Factors including friendly appearance and team player requirements were recorded and used in analyses to test models of customer and coworker discrimination. Employers could contact the candidate via mobile phone or email. Phone calls were not answered and instead missed calls were recorded and considered to be positive responses. Some firms sent a positive response directly by email, inviting the candidate to interview. Two days after a positive response through email, firms were notified that the application was withdrawn. If the company rejected the candidate, the observation was considered a negative response. Finally, if the firm did not respond at all, it was considered a rejection, in line with the approach used in literature (Bartos et al. 2016; Cahuc et al, 2017).

1.3 Results

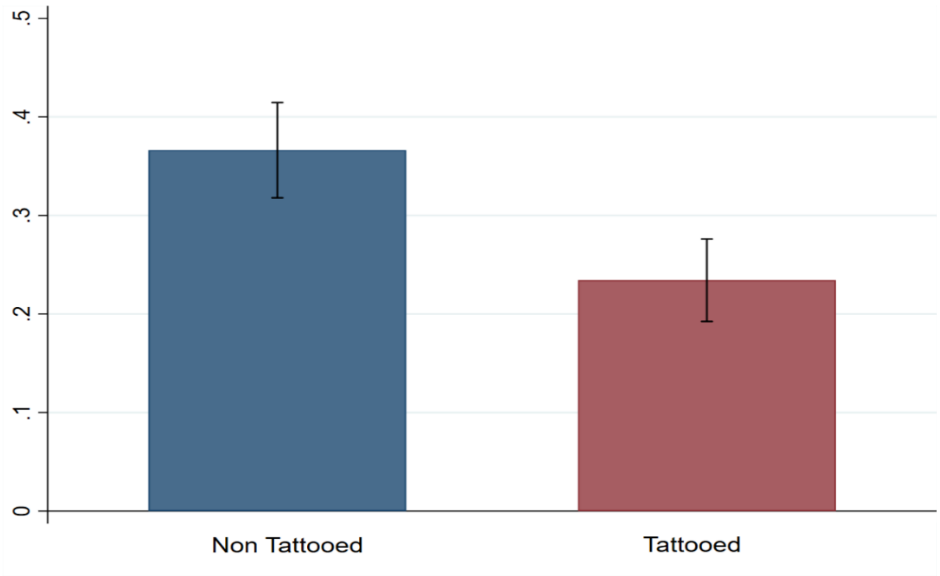
1.3.1 Descriptive Statistics and Balance Checks

I use Katuscak and Kraft (2013) study as a benchmark to conduct power calculations (and use a more conservative size effect of 10 pp difference between the treatment and control group than the 14 pp difference found by the authors), which suggests a sample size of approximately 800 observations. I responded to 782 job

[25] These words are (in German): bankkaufmann; kundenberater in bank; finanzberater, with the relevant English translations being bank clerk, customer advisor in a bank, and financial advisor.
[26] Data on the average earnings in each region and sector is available on the web portal of German Federal Statistical Office at: http://www.statistik-portal.de/Statistik-Portal/en/en_inhalt22.asp

advertisements in the experiment. Figure 1.2 shows the overall callback rate for tattooed and non-tattooed applicants. The difference is 13 percentage points (with *p-value=0.0001*). This difference translates into a reduction in callback rate of about 64% for applicants with visible tattoos. Thus, I observe a relatively high degree of discrimination against applicants with visible tattoos. The size of the effect in our study is similar to the difference between African American and White-sounding names reported in Bertrand et al. (2004), which is one of the largest effects compared to other correspondence studies in the field.

Figure 1.2: Callback rate across groups, %



Distribution of callbacks across different categories are summarized in Table 1.2. Column 1 shows the average callback for applicants with no tattoos and column 2 for applicants with tattoos. The percentage point difference in callback rates is shown in column 3, and the ratio of the callback rate in the control and the treatment group is shown in column 4.[27] The overall difference between callback rates for non-tattooed and tattooed applicants is 13% and is highly significant, even when adjusted with multiple hypothesis testing (List, Shaikh and Xu, 2016). As Table 1.2 shows, applicants with no tattoos always receive significantly more callbacks, and the difference ranges from 9pp (the difference for medium sized firms) to 19pp (for large firms). The only exception is when the position is posted for a rural area: in this case, applicants with no tattoos still have a higher callback rate, by 15pp; however, the difference is not statistically significant due to the small

[27] Numbers in columns 1 and 2 are rounded to the nearest decimal, while figures in column 4 are computed using the exact mean values.

number of observations. Our data show that the highest discrimination is observed in large firms (19pp difference, or a reduction of callback rate by about 98%).

Table 1.2: Distribution of callback across various groups

	Non-Tattooed	*Tattooed*	*Difference*	*Ratio*	*N*
	(1)	(2)	(3)	(4)	(5)
Overall	0.37 (0.48)	0.23 (0.42)	13***†††	1.56	782
Male	0.36 (0.48)	0.23 (0.42)	13***††	1.57	473
Female	0.37 (0.48)	0.24 (0.43)	13**	1.55	309
Front office	0.40 (0.48)	0.27 (0.42)	13**	1.48	335
Back office	0.34 (0.47)	0.21 (0.41)	13***††	1.63	447
West Germany	0.35 (0.48)	0.23 (0.42)	12***†††	1.54	647
East Germany	0.43 (0.49)	0.26 (0.44)	17*	1.64	135
Urban	0.36 (0.48)	0.23 (0.42)	13***†††	1.56	682
Rural	0.41 (0.49)	0.26 (0.44)	15	1.56	100
Medium firm	0.32 (0.47)	0.23 (0.42)	9***	1.44	358
Large firm	0.40 (0.49)	0.20 (0.40)	19***†	1.98	154

Notes: The table shows summary statistics of callbacks for our sample of 782 firms. Standard deviations are in parentheses beneath mean estimates. Column 3 shows the percentage difference in callback rates between the treatment (tattooed) and the control (non-tattooed) group. Column 4 reports the ratio of the first column to the second. * p<0.10; ** p<0.05; ***p<0.01. † - significance level with multiple hypothesis testing adjusted. Column 5 shows the number of observations in each subgroup.

In addition to lower callback rates, discrimination can take different forms. Table 1.3 shows the average reaction time (in working days) and the number of callbacks. As the table confirms, overall reaction time is significantly faster for candidates with no tattoos.[28] The effect mainly comes from the positive response, as firms call non-tattooed applicants significantly faster. This could indicate that tattooed candidates are close to the threshold of all applicants. Probably, firms first call candidates above the threshold and if they decline the offer, employers go down the list to choose an alternative. This would delay positive callbacks to applicants with tattoos. Because of the delay in positive

[28] Note that I have 148 missing observations for the variable "delay", when firms did not react at all. These observations are treated as rejection later in the regression analyses.

response, one might argue that if the experiment were run infinitely, at some point the treatment effect would disappear (the callback rate of applicants with tattoos will "catch up" to the callback rate of applicants without tattoos).

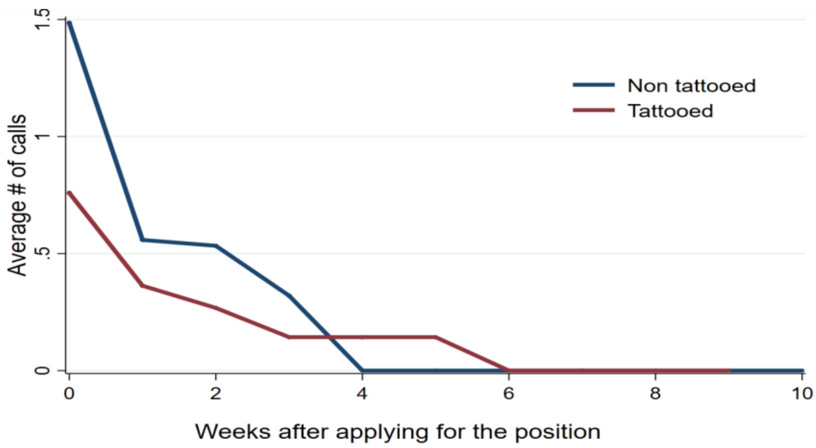
Table 1.3: Average reaction time in working days

Delay in response	*Non-Tattooed*	*Tattooed*	*Difference*
	(1)	(2)	(3)
All (*N=634)*	10.33 (13.75)	13.40 (21.45)	-3.07***
Positive response (*N=234)*	4.24 (5.16)	5.96 (6.92)	-1.72**
Rejection (*N=400)*	15.05 (16.26)	16.58 (24.46)	-1.53
Number of callback (*N=234)*	2.04 (1.66)	1.43 (1.36)	0.6***††

Notes: The table shows average reaction time in working days and number of callbacks across the treatment (tattooed) and the control (non-tattooed) groups. Standard deviations are in parentheses beneath mean estimates. * p<0.10; ** p<0.05; ***p<0.01. † - significance level with multiple hypothesis testing adjusted.

Figure 1.3 reflects Table 1.3 in the sense that the average number of calls is higher for non-tattooed applicants (blue line) compared to tattooed applicants (red line), and there is a delay in response to tattooed applicants. Further, as Figure 1.3 shows, firms stop calling back six weeks after receiving an application. Therefore, even if the experiment were conducted infinitely, the treatment effect would be maintained. In addition to the reaction time, Table 1.3 and Figure 1.3 confirm that employers call back applicants without tattoos more frequently; on average, about 1.4 times more. Because I do not answer the incoming calls from employers, they call several times in order to reach the applicant. This result indicates that they expend greater effort to reach applicants without tattoos.

Figure 1.3: Average callback across time



1.3.2 Estimation of the Linear Probability Model

In the previous section I showed that applicants with visible tattoos receive fewer callbacks than those without tattoos. To estimate the probability of a callback, the following linear probability model is estimated:

$$callback_i = \alpha_0 + \alpha_1 * Visible\ Tattoo_i + \beta_j * X_j + \varepsilon_i$$

where $callback_i = 1$ if applicant i gets a callback, and 0 otherwise; $Visible\ Tattoo_i = 1$ if the tattoo is attached to the picture of the applicant, and 0 otherwise; X_j represents the vector of covariates that includes firm/job characteristics and ε_i is the error term. Column 1 in Table 1.4 reports results of the simple model. In Column 2 I control for monthly fixed effects. Columns 3-5 expand the model (Column 2) to include dummy variables for gender, indicating whether the firm is international, and whether it is in an urban area, and their interaction with the treatment variable (the tattoo status), respectively. In these models, I also control for firms' characteristics including age, size and the number of jobs advertised by the firm. As column 3 shows, having a visible tattoo has the same effect for male and female candidates, as the interaction of the gender dummy with the treatment dummy is statistically insignificant. The data suggests that applying to an international firm further decreases the chances of a callback, as the coefficient is negative (column 4). However, the interaction is insignificant, meaning that tattooed applicants do not receive different treatment from international firms than applicants without tattoos. As for firms in urban areas, I see a similar pattern. Urban dummy is negative and the interaction with the treatment dummy is positive; however, it is not statistically significant.

Table 1.4: Estimates of the Linear Probability Model – Firm Characteristics

Dependent variable: Callback	(1)	(2)	(3)	(4)	(5)
Visible Tattoo	-0.13***††† (0.03)	-0.13***††† (0.03)	-0.13** (0.05)	-0.11***† (0.04)	-0.14 (0.09)
Male			-0.11** (0.05)		
Visible Tattoo * Male			-0.00 (0.07)		
International firm				-0.06 (0.05)	
Visible Tattoo * International firm				-0.05 (0.06)	
Urban area					-0.06 (0.08)
Visible Tattoo * Urban area					0.02 (0.10)
Constant	0.37***††† (0.02)	0.06 (0.15)	0.30** (0.15)	0.24 (0.16)	0.26* (0.15)
Monthly and regional dummies	*N*	*Y*	*Y*	*Y*	*Y*
Control variables	*N*	*N*	*Y*	*Y*	*Y*
R^2	0.02	0.09	0.13	0.12	0.12
N	782	782	782	782	782

Notes: Estimates of the linear probability model. Robust standard errors in parentheses. Columns 2-6 include monthly and regional dummies. In columns 3-6, I control for firm characteristics including age, size, number of job advertisements and whether the location of the job is in an urban area. * p<0.1, ** p<0.05, *** p<0.01. † - significance level with multiple hypothesis testing adjusted.

From Table 1.4, I conclude that none of the firm characteristics interacted with the treatment dummy have a significant effect on the probability of callbacks. I now consider whether job-specific characteristics have a heterogeneous effect on our applicants. As mentioned above, the job vacancies had different requirements, such as friendly appearance or teamwork skills, and concentrating on the banking sector allows us to examine discrimination in front- and back-office positions. These factors enable us to test the model of customer and coworker taste-based discrimination. In particular, if I observe that discrimination in applications for the front-office positions is greater compared to those for back-office positions, I could claim that tattooed applicants face customer taste-based discrimination. On the other hand, if in the positions where teamwork is required, I observe a greater level of discrimination compared to those without that particular requirement, I could argue that tattooed applicants suffer from coworker taste-based discrimination. As columns 2 and 4 of Table 1.5 show, I did not find evidence of customer or coworker taste-based discrimination. Even though the interaction terms in these columns are positive, they are both statistically insignificant. Column 3 of Table 1.5 reports that if the position has a requirement for a friendly appearance, it negatively affects the chance of getting a callback, though the effect is similar for applicants from both the treatment and control groups. I should note that after controlling for monthly fixed effects and firm characteristics, the treatment effect is roughly the same and highly significant.

Table 1.5: Estimates of the Linear Probability Model – Job requirements

Dependent variable: Callback	(1)	(2)	(3)	(4)
Visible Tattoo	-0.13***††† (0.03)	-0.13***††† (0.04)	-0.15***††† (0.04)	-0.13*** (0.05)
Front office		0.04 (0.05)		
Visible Tattoo * Front office		0.01 (0.07)		
Appearance requirement			-0.04 (0.05)	
Visible Tattoo * Appearance requirement			0.09 (0.07)	
Teamwork requirement				-0.08 (0.05)
Visible Tattoo * Teamwork requirement				0.07 (0.07)
Constant	0.06 (0.15)	0.24 (0.16)	0.28* (0.14)	0.55***† (0.19)
Monthly and regional dummies	*Y*	*Y*	*Y*	*Y*
Control variables	*N*	*Y*	*Y*	*Y*
R^2	0.09	0.13	0.12	0.12
N	782	782	782	782

Notes: Estimates of the linear probability model. Robust standard errors in parentheses. Columns 2-4 include monthly and regional dummies. In columns 2-4, I control for firm characteristics including age, size, number of job advertisements and whether the location of the job is in an urban area. * p<0.1, ** p<0.05, *** p<0.01. † - significance level with multiple hypothesis testing adjusted.

Finding no evidence of customer and coworker taste-based discrimination, I next examine the model of statistical discrimination. As mentioned above, tattooed individuals are linked to negative stigmas that makes them an "unwanted" group in the workplace. To provide positive signals about the personality of the applicants I used two methods: inclusion of contact details of a reference person in a fraction of applications, and inclusion of membership in the alpinist group in another fraction of applications.[29] Table 1.6 provides estimates of these two channels of statistical discrimination. Column 2 shows that inclusion of the contact details of the reference person decreases the probability of callback by 4%, though the effect is insignificant. The interaction term (of the treatment dummy and the reference signal) is also negligible in size and statistically insignificant. Column 3 in Table 1.6 shows that having a visible tattoo reduces the probability of callback by 16%. Similar to the reference signal, membership in the alpinists' association further decreases the probability of callback, though this is effect is negligible and statistically indistinguishable from zero. As the interaction term shows, being a member of the alpinists' association increases the probability of a callback by 8%; however, the effect is insignificant.

Table 1.6: Estimates of the Linear Probability Model – Channels of statistical discrimination

Dependent variable: Callback	(1)	(2)	(3)
Visible Tattoo	-0.13***††† (0.03)	-0.13***†† (0.04)	-0.15***††† (0.04)
Reference signal		-0.04 (0.05)	
Visible Tattoo * Reference signal		0.01 (0.07)	
Group membership			-0.00 (0.05)
Visible Tattoo * Group membership			0.08 (0.07)
Constant	0.06 (0.15)	0.28* (0.16)	0.25 (0.16)
Monthly and regional dummies	*Y*	*Y*	*Y*
Control variables	*N*	*Y*	*Y*
R^2	0.09	0.12	0.13
N	782	782	782

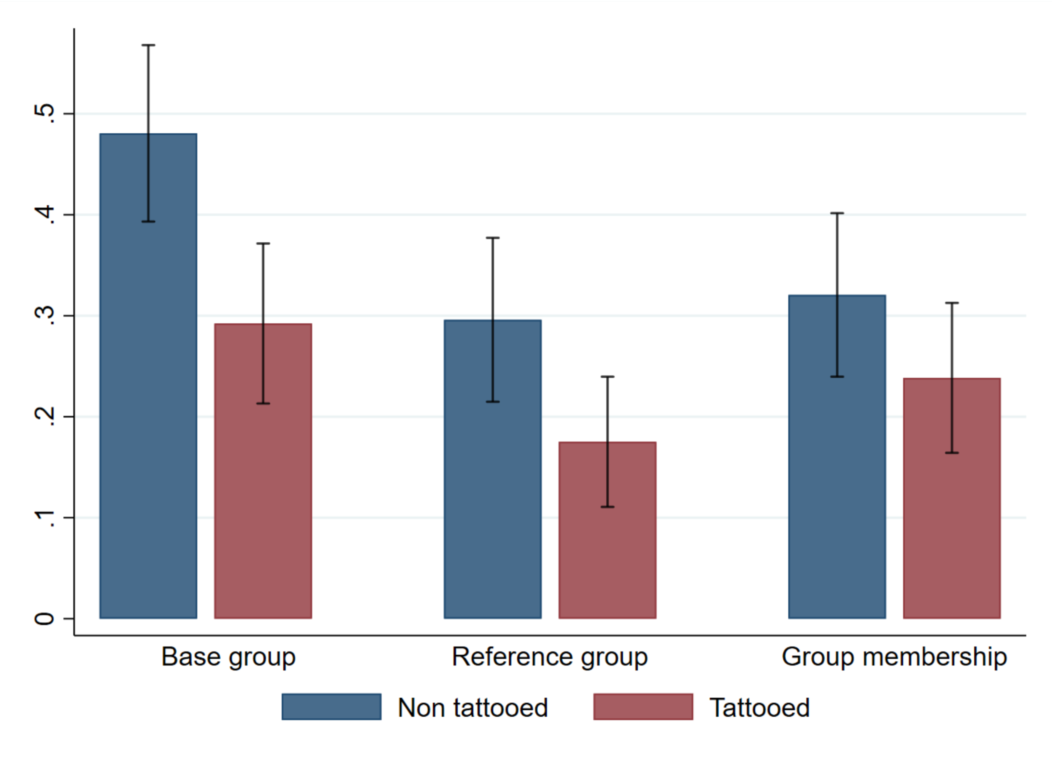
Notes: Estimates of the linear probability model. Robust standard errors in parentheses. All specifications control for monthly and regional dummies and firm characteristics including age, size, number of job advertisements and whether the location of the job is in an urban area. * p<0.1, ** p<0.05, *** p<0.01. † - significance level with multiple hypothesis testing adjusted.

Figure 1.4 considers the base group (no additional information), reference signal group, and alpinist membership signal groups separately. The difference in callback rates

[29] Usually, members of the alpinist group, Deutsche Alpenverein, are not risk-taking mountaineers, rather they are nature lovers who go on hiking trips. This should offset the negative stereotypes related to people with tattoos. Hence, I expected that membership would have a greater effect on applicants from the treatment group.

between applicants with and without tattoos in the base group is 0.19 with the *p-value=0.0018*; therefore, I can see that without providing any additional information regarding personality, applicants with tattoos are treated less favorably. In the reference signal group, the difference in callback rate is reduced to 0.12 and is still significant at the 5% level (*p-value=0.02*). On the other hand, in the alpinist membership signal group, the difference is no longer significant. Hence, using our channels I cannot reject the hypothesis that tattooed applicants face taste-based discrimination (possibly from HR managers).

Figure 1.4: Callback rate across groups – Channels of statistical discrimination



Therefore, unless I examine various other sources of statistical discrimination, I cannot present conclusive arguments about the nature of discrimination. However, my preferred interpretation is as follows. When I include the contact details of the reference person, the callback differential remains statistically significant. On the other hand, the alpinist group membership signal eliminates the difference. However, as shown in Figure 4, the callback rate declines for both tattooed and non-tattooed applicants when I include additional information regarding the personality of the candidate. The original purpose for including the additional information was to increase the callback rate for tattooed applicants. Figure 4 shows that it did not help. Therefore, in line with Gneezy et al. (2012), I claim that the discrimination against applicants with visible tattoos is motivated by animus. Specifically, it seems that HR managers in the banking sector have a distaste for workers with visible tattoos, resulting in differential treatment.[30]

1.4 Conclusion

In this paper I test whether applicants with visible tattoos are discriminated against in the banking labor market. As Akerlof and Kranton (2000) state, having a tattoo is an explicit signal of belonging to certain social categories.[31] Despite the growing popularity of tattoos, stigma still exists in society towards tattooed individuals. Because of these stereotypes, candidates with visible tattoos may be treated less favorably in the labor market, regardless of qualifications. Being a member of a particular social group does not necessarily predict individuals' productivity; thus, given equivalent qualifications, individuals with and without tattoos should have similar chances of employment. However, Fryer and Jackson (2008) showed that social categorization may lead to discriminatory behavior against minorities. This paper provides the first natural field experimental evidence to test whether applicants with visible tattoos are discriminated against in the labor market by looking at actual interview invitations.

To collect data, I conducted a natural field experiment in Germany. My data confirms that applicants with visible tattoos receive 64% less callbacks than those without tattoos. Further, I find that employers react positively significantly faster to candidates without tattoos, and they exert greater effort to contact the applicants (call back almost 2 times more). The design of the experiment allows me to test models of customer and coworker discrimination. However, I find no evidence that tattooed individuals face customer and/or coworker taste-based discrimination. Moreover, I use two channels to test whether tattooed applicants suffer from statistical discrimination. Even though I have mixed evidence regarding statistical discrimination, I suggest that the discrimination found in this study might come from animus, particularly from HR managers. This is in line with the argument made by Gneezy et al. (2012), in which the authors argue that when the object of discrimination is perceived as a choice made by the individual, discrimination is taste based. However, I should note that various other sources of

[30] I have analyzed our data using a probit estimation method; however, qualitative results remain unchanged.
[31] Even though I did not use tattoos that have a clear connection to some specific social groups, tattoos are generally viewed as a hindrance factor in the labor market, at least in some occupations.

statistical discrimination should be tested to eliminate this form of discrimination and argue that visibly tattooed applicants face taste-based discrimination.

In the context of models by Akerlof et al. (2000) and Fryer et al. (2008), I find that job applicants are often punished for expressing their identity or belonging to a certain social group. My results show how important identity is and what consequences job applicants may face when they express it. This might be problematic from the firms' perspective, as they can lose talent as a result of discrimination. Applicants, on the one hand, may suffer from discrimination, as they need more time and resources to find a job. On the other hand, given that applicants have information about discrimination against tattooed individuals, those who want to work in a more "liberal" environment may choose to signal the employer by using tattoos to screen themselves from discriminatory firms. Thus, in addition to its contribution to the literature, the results of this paper may be useful for members of the general public.

Having a tattoo potentially has a heterogeneous effect across sectors. Therefore, in order for the results to have a higher degree of external validity, a broad range of occupations should be considered. There are other possible extensions of the paper. In this paper I use "neutral" tribal tattoos; however, the content of the tattoo may have a different effect. For example, a tattoo that can elicit a positive emotion may not be viewed as negatively as a tattoo that elicits a negative emotion. In addition, the size of the tattoo may have a different effect on employment chances. Thus, more research in this area is needed to find out why/if HR managers "punish" individuals for expressing themselves.

2. Tattoos and Employment: Examining the Impact of Visible Tattoos in the IT Sector [32]

2.1 Introduction

Previous empirical research has documented unfavorable treatment of applicants with visible tattoos in various settings (see Jibuti, 2018; Dolaec and Stein, 2013; Dillingh et al., 2019; Timming et al., 2017). Jibuti (2018) conducted a field experiment in Germany and found that applicants with a visible tattoo receive about 64% fewer callbacks, an effect size comparable to that found by Bertrand and Mullainathan (2004). However, the paper focused on the banking sector, where, in addition to skills, appearance is also an important factor affecting the employment chances. Therefore, one might a priori expect discrimination against tattooed applicants in the banking sector and argue that it may be sector specific. In the current study, I extend the analysis to the more skill-intensive IT sector to examine whether this form of discrimination persists in completely different settings. Moreover, IT sector employees usually have minimal face-to-face interaction with customers, reducing room for appearance-related potential bias in hiring. Additionally, unlike Jibuti (2018), I introduce variation in the skill level of applicants to test for statistical discrimination, as suggested by Neumark (2012). I conducted a correspondence study in Germany to collect and analyze data and found that applicants with visible tattoos are 25% less likely to receive a callback, an effect that remains significant after controlling for company/position related characteristics and monthly and regional dummies. After failing to find evidence for tattooed applicants facing statistical discrimination, I suggest that hiring bias may be driven by taste-based discrimination or beliefs about unobserved personal characteristics. The results of our study, combined with Jibuti's (2018) findings, would provide a more comprehensive understanding of discrimination against visibly tattooed applicants.

Despite the increasing popularity of tattoos (Walzer and Sanjurjo, 2015), negative stereotypes are still associated with tattooed individuals (Antonellis and Silsbee, 2018)[33], which may drive an unfavorable treatment against these applicants in various markets.

[32] I am grateful to Randall Filer, Andreas Menzel and Michal Bauer for their valuable feedback on the earlier version of the paper. The CERGE-EI dissertation committee approved conducting the experiment during the dissertation workshop proposal in June 2015. All remaining errors are mine.

[33] This is particularly true for individuals with visible (face/neck) tattoos, as it is considered to be a more extreme form of body art.

However, the psychology literature finds very little difference in the personality traits of individuals with or without tattoos (Swami et al., 2016). In economics, only a handful of papers have explored the same issue. In an incentivized experiment Ruffle and Wilson (2018) find that the behavior of tattooed and non-tattooed individuals did not differ in honesty-decision tasks. The result was robust in low or high-stake tasks and the number and placement of tattoos (visible or not) did not affect findings. On the other hand, another experiment by the same authors shows that tattooed individuals, particularly those with visible tattoos, are more short-sighted and impulsive compared to people without tattoos (Ruffle and Wilson, 2019). The authors also find that tattooed individuals substantially underestimate the potential negative impact tattoos may have in the labor market.[34] Ruffle and Wilson (2019) argue that if short-sightedness and impulsivity are the reason for discrimination, they would expect a high level of discrimination in professions where patience and planning skills are required, and a low level – or absence of – discrimination where quick decision-making takes place. Our paper's results, combined with those of Jibuti (2018), can thus provide empirical evidence for this hypothesis.

Evidence suggests that differences in personal characteristics of tattooed and non-tattooed individuals are disappearing as tattoos are becoming more widespread (Swami et al. 2016). However, visibly tattooed applicants continue to face unfavorable treatment in many different settings. In addition to Jibuti (2018), other empirical research has also confirmed negative bias against tattooed applicants. Interesting work related to our paper experimentally tests whether minority (black) sellers face discrimination in the online market (Dolaec and Stein, 2013). The authors posted online advertisements offering iPods on websites throughout the US. To signal the race of the seller, the authors use photographs of a dark- or light-skinned hand holding the item. In addition to signaling race, a photograph of a light-skinned hand with a wrist tattoo was also used as a social signal. The authors propose that this latter group "can serve as a 'suspicious' white control group" (Dolaec and Stein, 2013, p. 2) and they may face the same level of discrimination as black sellers. The results of the experiment suggest that advertisements with a black or tattooed hand in the photograph receive significantly worse treatment than those with white hands. This is true for all outcome measures that the authors observe, including responses, offers, amounts offered and trust. Dolaec and Stein (2013) find some evidence

[34] This may be because the correlation between tattoos and potential employment outcomes is a relatively underexplored question in economics research; thus, individuals may make uninformed decisions about where to place a tattoo.

of black and tattooed individuals suffering from statistical discrimination (Phelps, 1972). While the paper focuses on customer discrimination, the authors note that it may cause other forms of discrimination: if customers do not wish to buy products from tattooed sellers, retailers will avoid hiring tattooed workers. Therefore, discrimination in the advertisement market may translate into unfavorable treatment in the labor market.

While the papers described above find evidence of a potential negative correlation between tattoos and market outcomes, some recent research was unable to establish a strong relationship. Dillingh, Kooreman and Potters (2019) use panel data from various surveys of Dutch individuals to examine the correlation between tattoo status and several outcome measures, including income in the relevant year and employment status. The authors find that irrespective of gender, having a tattoo is negatively correlated with monthly income, although the effect is not statistically significant. On the other hand, having a visible tattoo increases the likelihood that a person is unemployed. Using the survey data from Mechanical Turk (MTurk) participants, French et al. (2019) study whether tattoos adversely affect labor market outcomes. After controlling for conventional socio-demographic variables (age, gender, race, education, etc.), the authors fail to find evidence of a significant correlation between tattoos and employment or earnings. These results could suggest that in some societies (the Netherlands and France in this case) tattoos are no longer a burden after controlling for demographic variables. However, they also show that tattooed individuals have lower education and less favorable physical/mental health conditions (Dillingh et al., 2019). The advantage of our experimental design is that I can compare outcomes of applicants with exactly the same qualifications, but with or without tattoos, which is not feasible in these observational studies. Therefore, while those empirical studies are informative for our analysis, I believe that this experiment is a more robust research method for detecting potential discrimination against tattooed applicants.

Correspondence studies are extensively used in the discrimination literature[35], with most papers focusing on the object of the discrimination that is exogenous for the person, such as age (Neumark et al., 2019), race (Bertrand and Mullainathan, 2004; Bartoš et al., 2016; Nüß and Penny, 2019), gender (Baert et al., 2016a) and physical attractiveness (Ruffle and Shtudiner, 2015). Recently, a small, but growing stream of literature has emerged which examines characteristics that are a choice of an individual.

[35] Neumark (2018) and Baert (2018) provide an excellent overview of research using correspondence testing method to examine labor market discrimination.

Other objects of potential discrimination include religion (Weichselbaumer, 2019), military service or affiliation (Baert and Balcaen, 2013), marital status (Arceo-Gomez and Campos-Vazques, 2014), motherhood (Becker et al. 2019) and job changes (Cohn et al. 2017).

Cohn et al. (2017) hypothesize that frequent job changes may signal poor non-cognitive skills (reliability, trustworthiness, loyalty, and self-control (Lindqvist and Vestman, 2011)), which in turn may translate into discrimination against workers who switch jobs frequently. In multiple experiments, the authors find that individuals who switch jobs more often have a lower chance of employment. Frequent job changes, particularly between industries, may decrease worker productivity, which Cohn et.al (2017) also find. Thus, the discrimination found in Cohn et al. (2017) may be the result of lower perceived productivity, which is consistent with statistical discrimination. Tattoos, on the other hand, do not necessarily signal lower productivity, and having a visible tattoo should not therefore diminish employment chances due to lower perceived productivity, unless HR managers have mistaken beliefs about the productivity of tattooed applicants. By manipulating information about the productivity of applicants, I can test whether differences in beliefs about productivity drive discrimination.

I conducted a correspondence study in Germany to collect data. Since the inclusion of a photo is still a common practice in Germany and the labor market is large, it provides an ideal setting for the study. As proposed by Neumark (2012), I randomly varied the level of qualifications of applicants (high vs. low skilled), to address Heckman and Siegel's (1993) critique of potential bias in an estimate due to the non-linearity of the outcome variable in applicant's productivity. I expect high-skilled applicants to have a higher callback rate than their low-skilled counterparts, irrespective of tattoo status, as empirical evidence suggests earnings premia for advanced degree graduates (Altonji and Zhong, 2020). Varying the skill level of applicants will also help in detecting and potentially reducing statistical discrimination (Bertrand and Duflo, 2017). If the level of discrimination is reduced (or eliminated) for highly skilled applicants, it could be interpreted as evidence of statistical discrimination. Therefore, I hypothesize that level of discrimination against tattooed applicants should be reduced for high-skilled candidates. In addition to the skill level, I also randomly vary the gender of the applicant and examine how it affects the callback rate. Empirical studies in the social psychology literature have found negative attitudes towards women with (visible) tattoos (Guéguen, 2013). Timming et al. (2017) also show that women with tribal-themed tattoos are perceived to be less trustworthy compared to men. Hence, I conjecture that tattooed female applicants face a higher level of discrimination compared to tattooed male applicants. Additionally, since the IT sector is male dominated[36], I expect a lower callback rate for the female applicants in the experiment, irrespective of tattoo status.

My findings indicate that visibly tattooed applicants have lower employment chances even in the skill-intensive IT sector. Even after controlling for additional variables, the impact of tattoos remains statistically significant. Contrary to expectations, female applicants have a higher chance of callback, although the effect of tattoos is homogeneous across gender. Moreover, increasing the skill level of applicants does not help to reduce the unfavorable treatment of tattooed applicants. I find no evidence of statistical discrimination, which is in line with Weichselbaumer (2019), who argues that the level of information provided in job applications in Germany leaves little room for statistical discrimination. Therefore, I interpret our results as evidence suggestive of taste-based discrimination, which is in line with my previous finding (Jibuti, 2018), which suggests that tattooed applicants face taste-based discrimination in the banking sector.

The next section describes the experimental design and data description. Section 3 summarizes key results and discusses robustness checks. Section 4 concludes.

2.2 Experimental Design

The experimental design builds on Jibuti (2018) and uses professional fictitious applications to apply for actual vacancies from the IT sector.[37] I opted to use photos of only average-looking female and male applicants, as no evidence of the impact of beauty on a callback decision was found in Jibuti (2018). Additionally, I randomly varied (orthogonally to tattoo status) the skill level of applicants to address the criticism of correspondence studies by Heckman and Siegel (1993). These researchers argue that correspondence studies can generate biased estimates due to the non-linearity of the outcome variable in productivity. Although researchers can control and match some

[36] According to the German Federal Statistical Office data (available at the following link), only about 32% of all employers in the Information and Communication sector were female in 2019. More detailed data from the US Bureau of Labor Statistics suggest that in the occupation "Computer programmers" women constitute about 20% of all employment.
[37] Sample resumes can be found in Appendix B

productivity measures of different groups (with or without tattoos, in our case) in fictitious applications, companies also consider unobserved productivity characteristics during their hiring decision. According to Heckman and Siegelman (1993), employers will hire applicants if the sum of observed and unobserved productivity characteristics is sufficiently high. Even if unobserved productivity is the same across groups, the variance of unobserved productivity may be different. Assume, for example, that an employer believes that the variance of unobserved productivity factors is higher for applicants with visible tattoos. In this case, the employer would correctly conclude that applicants with visible tattoos have a greater chance of having a higher sum of observed and unobserved productivity measures. Thus, applicants with visible tattoos will be hired more frequently, consistent with the statistical discrimination models (Aigner and Cain, 1977).

Neumark (2012) suggested that by varying the observable productivity measures researchers would be able to overcome Heckman and Siegel's (1993) criticism and recover an unbiased estimate of discrimination. Therefore, applications are designed to have two levels of qualifications, representing low and high levels of skills. I differentiate high- and low-skilled applicants in several productivity measures. The high-skilled applicant has better educational outcomes, including a higher high school/university Grade Point Average (GPA) and attainment of a higher education degree in computer science (master rather than bachelor). Moreover, the high skilled applicant has about 2 years more experience (at a well-known multinational company, rather than a less-well-known German company) and knowledge of more programming languages/software than the low-skilled applicant. According to the HR application agency that designed the CVs, our low-skilled applicant can be considered to be an "average candidate" and the high-skilled applicant an "above average candidate" on the German IT labor market at the time of running the experiment.

2.2.1 Process and Data Description

The experiment was conducted between March 2018 and May 2020. The position of Software Developer (in German: Software-Entwickler/in) was searched for on one of the largest online job boards in Germany (Stepstone.de). Applications were combined into a single PDF file and were submitted via email (or through the online application system) in response to job offers for which our applicants satisfied minimum requirements. When available, detailed information about the company was recorded at the time of applying for the position, including contact details (phone number and/or email address), number of employers, age of the company, and location and gender of the HR contact person. Some of this information was helpful in identifying the company if they responded, while the remainder was used in the analysis. Some employers requested (via email) university/employer certificates along with the application. However, in order to avoid having to create fictitious university degrees, I opted not to create those documents and informed a company that, if invited, the candidate was happy to bring the requested documents for the personal interview. About 16% (with s.d.=0.37) of firms required additional certificates and requests did not differ across the control and treatment groups. To indicate the desired salary in the application, I matched average yearly earnings by regions to the location of the workplace, based on data from the German Federal Statistical Office.

After the application was submitted, I recorded the response of employers and matched them to the company in our database. Firms usually called on the phone and/or sent an email. To minimize the cost to employers I did not answer the phone calls. Callback and/or an email regarding an invitation for a personal/phone interview was considered to be a positive response. After receiving an email from the company, I declined the offer within a few days. The negative response also had two forms: either the company rejected an applicant via email, or they did not respond to the application at all. Both cases were treated as a rejection of the applicant.

In total, I submitted 800 applications. Table 2.1 shows descriptive statistics and balance checks in our dataset. Since our experiment used a between-firm design, I wanted, by using balanced checks, to ensure that companies/positions receiving applications from tattooed and non-tattooed applicants did not differ significantly in terms of measurable characteristics. Table 1 shows that I fail to reject the null hypothesis of the equal mean for all the characteristics that I observe. This confirms that firm or position characteristics are similar in the treatment and control group; hence, any difference in the callback rate should not be associated with those characteristics. The smallest p-value, 0.14, is observed for the dummy variable for female HR contact: the HR contact person was female in 69% of cases for non-tattooed applicants and 63% of cases for tattooed applicants. I don't believe that this may drive our results, as there is no evidence of females having different attitudes towards tattooed individuals (Zestcott et al. 2017).

Table 2.1: Summary statistics and balance check

Firm/job characteristic	*Overall mean*	*Obs.*	Dep. variable	
			Estimate	*p-value*
Firm size	1324 (2364.82)	800	0.00 (0.00)	0.79
Firm age	37 (33.79)	800	-0.00 (0.00)	0.66
International firm	0.45 (0.49)	800	0.04 (0.04)	0.23
Number of ads	12 (20.51)	800	-0.00 (0.00)	0.75
Female HR contact	0.66 (0.47)	800	-0.06 (0.04)	0.15
High skilled applicant	0.47 (0.49)	800	-0.01 (0.04)	0.82
Teamwork requirement	0.44 (0.49)	800	0.01 (0.04)	0.87
# of required programs/software	6 (3.44)	800	0.00 (0.01)	0.86
"Senior" in the position title	0.09 (0.29)	800	-0.00 (0.06)	0.99
Submit email vs. online	0.65 (0.48)	800	-0.02 (0.04)	0.64
Urban area	0.77 (0.42)	800	0.01 (0.04)	0.80
Joint significance test				0.95

Notes: The table shows the balance check of the data. The first column shows means of firm/position characteristics. Standard deviations are in parenthesis beneath mean estimates. The third column shows estimates of the treatment dummy on the corresponding characteristic. The last column shows p-values of the estimates.

2.3 Results

2.3.1 Descriptive Summary

Figure 21 shows bar charts with confidence intervals of overall callback rates for applicants with and without tattoos. The difference between the two groups is 9 percentage points (p-value=0.009), or a 33% higher callback rate for non-tattooed applicants. Considering the settings of the study, the size of the effect is large, although it is lower than the effect found in Jibuti (2018), which was 56%.

Figure 2.1: Callback rates across groups, %

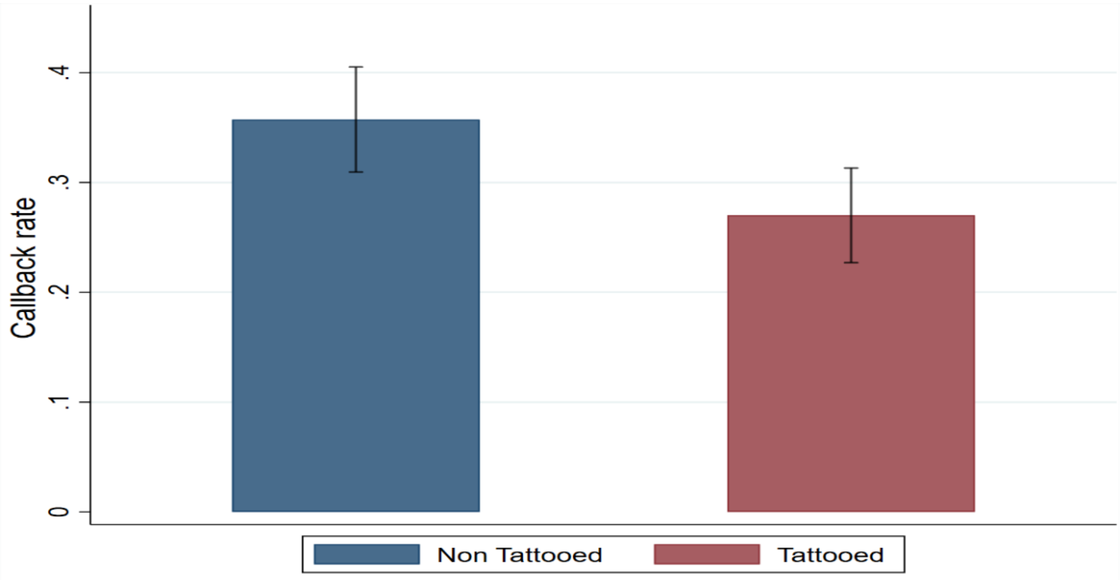


Table 2.2 shows the overall callback difference between the non-tattooed and tattooed applicants for different subgroups in the sample. Average callback rates for non-tattooed and tattooed applicants in the respective subgroups are shown in columns 1 and 2. Column 3 shows the percentage point difference between columns 1 and 2, while the last column shows the ratio of the two. As shown in the table, the overall difference in callback rates between non-tattooed and tattooed applicants is 9 percentage points and is significant at the 1% level. The difference is very similar within the sub-groups of male and female applicants; the slightly larger difference for female applicants is not significantly different from that for male applicants. The effects of tattoos seem larger for high-skilled applicants, with a significant gap of 11 pp in call-back-rates, while the gap is 7 pp and not statistically significant for low-skilled applicants. However, I cannot reject the hypothesis that the gap is the same for high- and low skilled applicants.

Table 2.2: Distribution of callback across the treatment (tattooed) and control (non-tattooed) group

	Non-Tattooed	*Tattooed*	*Difference*	*Ratio*	*N*
	(1)	(2)	(3)	(4)	(5)
Overall	0.36 (0.48)	0.27 (0.44)	9***	1.32	800
Male	0.32 (0.47)	0.24 (0.43)	8*	1.32	403
Female	0.40 (0.49)	0.29 (0.46)	10**	1.34	397
High skilled appl	0.38 (0.49)	0.28 (0.45)	11**	1.37	373
Low skilled appl	0.33 (0.47)	0.26 (0.44)	7	1.28	427
West Germany	0.37 (0.48)	0.27 (0.44)	10**	1.37	681
East Germany	0.29 (0.46)	0.27 (0.44)	2	1.09	119
Urban	0.35 (0.48)	0.27 (0.45)	8**	1.29	618
Rural	0.37 (0.48)	0.25 (0.44)	12*	1.47	182
Small firm	0.44 (0.49)	0.26 (0.44)	18***	1.71	207
Medium firm	0.33 (0.47)	0.28 (0.45)	6	1.20	379
Large firm	0.31 (0.47)	0.26 (0.44)	5	1.19	211

Notes: The table shows summary statistics of callbacks for our sample of 800 firms. Standard deviations are in parentheses beneath mean estimates. Column 3 shows the percentage difference in callback rates between the treatment (tattooed) and the control (non-tattooed) group. Column 4 reports the ratio of the first column to the second. * p<0.10; ** p<0.05; ***p<0.01. Column 5 shows the number of observations in each subgroup.

Another interesting result is apparent when comparing callback differences across geographies: firms discriminate more in western Germany, while the difference in callback rate is relatively small and insignificant for eastern Germany. One possible explanation might be that over 85% of our sample is collected from western Germany, thus limiting the power of the test for the eastern region of the country. On the other hand, firms in the western region potentially have more applicants, and thus may discriminate more, while firms in the eastern part "cannot afford" to discriminate as they have a limited number of applicants.[38] Interestingly, I observe larger discrimination in rural areas,[39] even though only about 23% of our sample is collected in such areas. As for the size of the

[38] According to the report by the Ministry of Economic Affairs and Energy (2019), the unemployment rate was 6.9% in east Germany in 2018, while it was 4.8% in the western part of the country.
[39] Urban dummy is defined as a location with more than 100 000 inhabitants.

firm,[40] our data suggest that small firms discriminate most, and the difference between callback rates for tattooed and non-tattooed applicants is highly significant. I did not find evidence for other forms of discrimination, including delay in response or the number of callbacks.

Jibuti (2018) notes that employers may also differentiate how quickly they respond to applications received from different groups. Table 2.3 below shows the summary of reaction time across treatment/control and gender groups. Overall, tattooed applicants in our data receive a response (positive or negative) 2 days later than do non-tattooed applicants, though this difference is not statistically significant (row 1 in Table 2.3). Therefore, the data does not suggest that tattoed applicants experience significant delays in employer responses, compared to non-tattooed applicants. The table also shows that there is no statistically significant difference in delays for applicants based on gender: female and male applicants experience similar delays in responses irrespective of having a visible tattoo or not (rows 2 and 3 in Table 2.3). However, within the male group, tattooed applicants have a larger delay in response (4 days) compared to the female group (0 days), suggesting that employers are punishing male applicants more than female applicants for having a visible tattoo. As the data suggests, employers start to respond to non-tattooed male applicants after they have already responded to tattooed female applicants. Moreover, if we compare the response delay between female and male tattooed applicants (12 and 17 days, respectively), the difference is marginally significant (not displayed in the table). This result is in contrast to previous work, which found more negative attitudes against women than men with tattoos(Guéguen, 2013; Timming et al. 2017). One possible explanation is that, as a male-dominated sector, employers in the IT sector are more willing to engage with female applicants. Indeed, column 3 in Table 2.4 (section 2.3.2) shows that male applicants have, on average, an 8% lower chance of a callback. The last row in Table 2.3 confirms that employers call tattooed applicants significantly less often, though the size of the effect is very small.

[40] Firm is defined to be small if they employ less than 100 workers; medium, if their workforce is between 100 and 1000; big, if they employ more than 1000 workers.

Table 2.3: Average reaction time in working days

Delay in response	*Non-Tattooed*	*Tattooed*	*Difference*
All (*N=573)*	12.46 (14.39)	14.51 (22.47)	-2.04
Female (*N=287)*	11.95 (13.36)	12.5 (10.50)	-0.09
Male (*N=286)*	12.98 (15.39)	16.98 (29.89)	-4.01
Number of callback (*N=573)*	1.25 (1.75)	0.86 (1.48)	0.39***†

Notes: The table shows average reaction time in working days and number of callbacks across the treatment (tattooed) and the control (non-tattooed) and gender groups. Standard deviations are in parentheses beneath mean estimates. * p<0.10; ** p<0.05; ***p<0.01. † - significance level with multiple hypothesis testing adjusted.

As Bertrand and Duflo (2017) note, one of the limitations of correspondence studies is that the outcome variable is a callback dummy. Indeed, only a handful of studies have also considered outcomes other than callback: Weichselbaumer (2017) examined response times and found that ethnic minority applicants receive an interview invitation significantly later than do ethnic majority applicants. Additionally, ethnic minority applicants are rejected significantly faster than ethnic majority applicants. Bartos et al (2016) analyze how the number of pieces of information acquired or the number of attempts to open a resume differs across ethnic groups. The authors show that more information is acquired for ethnic minorities and that a name that is recognizable as a minority reduces the effort to open the resume. Dolaec and Stein (2013) also analyze different outcome measures, among them the number of non-scam responses to their advertisements. They find that black and tattooed sellers receive significantly fewer responses than whites. Thus, my study uses a range of outcomes, some of which have not yet been analyzed extensively in the literature, such as response time.

2.3.2 Linear Probability Model

This section estimates a linear probability model of the callback rate, including interactions with a number of control variables. The dependent variable is a dummy for a callback and the main independent variable is tattoo status. The results using Probit and Logit models are qualitatively the same as the OLS estimates. Table 2.3 reports the results of the linear probability model with various specifications. Column 1 shows a simple regression without any controls. Columns 3-7 test whether gender, skill level, workplace location and firm size have a heterogeneous impact on the callback rate of tattooed applicants. In addition to these main variables of interest, the models in Columns 2-7 include monthly and regional fixed effects and controls for various firm/job characteristics, including firm age and size (measured by number of employees), number of job advertisements, number of required programs/software, gender of the HR contact person, whether the position requires teamwork, whether the location of the job is in an urban area and whether the firm is international.

The first column of Table 2.4 suggests that applicants with visible tattoos are about 9 percentage points less likely to get a callback. The effect is statistically significant and is lower in magnitude than the effect found in Jibuti (2018). When monthly and regional fixed effects are included in the regression, the treatment effect is slightly lower (-0.07, *p-value=0.04*), though it remains significant (column 2). Column 3 extends the model to investigate how gender subgroups are affected by a visible tattoo. The treatment dummy variable remains negative and marginally statistically significant. Contrary to expectations, the results show that male applicants are about 8% less likely to receive a callback than females, and the estimate is statistically significant. On the other hand, the interaction of males and the treatment dummy is positive, though statistically indistinguishable from zero (Column 3). Thus, I cannot reject the null hypothesis of an equal treatment effect for our male and female applicants. Our data suggest that the gender of the applicant is largely irrelevant when the potential candidate has a visible tattoo.

Column 4 examines how the level of skills affects the treatment of applicants with visible tattoos. In line with expectations, high-skilled applicants are about 5% more likely to get a callback, though the effect is insignificant. The interaction term of high skill and treatment dummy is negative, but also insignificant, suggesting that the impact of tattoos does not differ by skill levels. As described in the experimental design, the level of skills was controlled to test whether tattooed applicants face statistical discrimination. If having a higher skill level would reduce or eliminate discrimination against tattooed applicants, it would be evidence of statistical discrimination. However, my results do not confirm this hypothesis; hence, I suggest that tattooed applicants may face taste-based discrimination. Alternatively, HR managers may discriminate based on their beliefs about unobserved personal characteristics of tattooed candidates, such as patience, kindness and acceptance of rules.[41]

[41] These beliefs itself may be related to (usually incorrect) prejudices associated with tattooed people.

Table 2.4: Estimates of the Linear Probability Model

Dependent variable: Callback	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Visible Tattoo	-0.09*** (0.03)	-0.07** (0.03)	-0.09* (0.05)	-0.06 (0.05)	-0.01 (0.09)	-0.10 (0.07)	-0.03 (0.06)
Male			-0.08* (0.05)				
Visible Tattoo * Male			0.04 (0.07)				
High skill				0.05 (0.05)			
Visible Tattoo * High skill				-0.04 (0.07)			
Visible Tattoo * West Germany					-0.07 (0.09)		
Urban area						-0.02 (0.06)	
Visible Tattoo * Urban area						0.03 (0.08)	
Small firm							0.10 (0.09)
Visible Tattoo * Small firm							-0.14 (0.09)
Medium firm							-0.01 (0.07)
Visible Tattoo * Medium firm							-0.02 (0.08)
Constant	0.36*** (0.02)	0.68*** (0.24)	0.74*** (0.24)	0.65*** (0.23)	0.64*** (0.27)	0.69*** (0.24)	0.67*** (0.24)
Monthly and regional dummies	*N*	*Y*	*Y*	*Y*	*Y*	*Y*	*Y*
Control variables	*N*	*Y*	*Y*	*Y*	*Y*	*Y*	*Y*
R^2	0.009	0.068	0.073	0.070	0.069	0.068	0.074
N	800	800	800	800	800	800	800

Notes: Estimates of the linear probability model. Robust standard errors in parentheses. Columns 2-7 include monthly and regional dummies in the regression. In columns 2-7, I control for firm characteristics including age, size, \# of job advertisements, \# of required programs, gender of HR contact, whether the position includes "senior" in the title or requires teamwork, whether the location of the job is in an urban area and whether the firm is international. * p<0.1, ** p<0.05, *** p<0.01.

Columns 5 and 6 examine how the treatment of tattooed applicants differs across geographical locations. Carlsson and Rooth (2012) argue that discrimination against minorities is higher in regions where a more negative attitude exists towards minority groups. Even though no evidence exists of differential attitudes towards tattooed individuals across different locations in Germany, I hypothesize that tattooed individuals may be evaluated differently in eastern than western Germany, or in rural compared to urban areas, leading to a different level of discrimination against tattooed applicants in those locations. Column 5 suggests that applying for the position from western Germany increases the chance of callback by 3%, although the estimate is not significant. Most importantly, tattooed applicants are not treated differently in the western or eastern Germany, as the interaction term is statistically insignificant. Similarly, the interaction term of tattoo status and the urban dummy is insignificant (column 6), implying that visible tattoos do not have a significantly different impact in rural or urban areas. To sum up, the treatment effect in both geographical locations is insignificant, albeit large. This may suggest that the gaps are in fact different, though our data lacks the statistical power to reject the null hypothesis of equal treatment.

Another job-related characteristic that may also influence the decision is firm size. Kaas and Manger (2012) claim that the hiring process in small firms is less standardized, which would leave more room for potential bias in hiring. Therefore, one would expect higher discrimination against minorities in small firms. I examine this hypothesis in the regression analysis and column 7 shows the estimation results. Applying for a position in a small firm (less than 100 employees) increases callback probability by 10% while applying to medium-sized firms (100 – 1000 employees) reduces the probability of callback by 2%. However, both estimates are statistically insignificant. Small and medium firms seem to discriminate against tattooed applicants more, as evidenced by the negative interaction terms in Column 7. Although both interaction terms are statistically insignificant, the treatment effect is large for small firms, with our sample lacking the statistical power to reject the null hypothesis of equal treatment.[42]

To summarize the results, the core finding of this paper is a strong and robust impact of visible tattoos on the probability of callback regarding a job application: a significant 9 percentage points reduction for applicants with tattoos. This implies that applicants with visible tattoos have about a 25% lower chance of getting a callback in the IT industry in Germany. Neither firm/position-related characteristics nor monthly or regional fixed effects mitigate this result. Moreover, I did not find evidence of statistical discrimination as high- and low-skilled tattooed applicants were treated, on average,

[42] I also examined whether other job or company-related characteristics, including age/size of the company, whether the company is multi-national, and whether the gender of the HR contact person would affect discrimination against tattooed applicants. However, similarly to the results reported in Table 3, none of those characteristics had an impact on the treatment of tattooed applicants, as the interaction terms remain statistically insignificant.

similarly. Tattoo impact also did not differ by gender (male/female), geographical location (eastern/western Germany or urban/rural area) and firm size (small vs medium and large), as all the relevant interaction terms were statistically insignificant in the regression analysis. However, I should emphasize that the treatment effect was insignificant in more than half of the models in Table 2.4, and hence I lack the statistical power to conclusively suggest that the impact of having a tattoo is homogeneous across those subgroups.[43] Our results may indicate that applicants with visible tattoos face taste-based discrimination: besides failing to find evidence of statistical discrimination, treatment of tattooed applicants did not differ across the groups mentioned above. If indeed true, this suggests that employers in the IT sector in Germany have rather uniformly negative attitudes towards tattooed applicants, irrespective of gender, skill level or any other firm/position related characteristics. Similar results were obtained in Jibuti (2018) in the banking industry, though the magnitude of discrimination was larger than in the IT sector. If these results generalize to other sectors, HR managers in occupations even with limited face-to-face customer interaction may be reluctant to hire applicants with visible tattoos.

2.4 Conclusion

In this paper I examine whether applicants with a visible neck tattoo are discriminated against in the labor market in the IT industry. Tattoos are no longer related to specific groups (sailors, gang members, etc.), but are widespread within diverse segments of society (Walzer and Sanjurjo, 2015). Moreover, empirical evidence suggests that there is very little difference in personal characteristics between tattooed and non-tattooed individuals (Swami et al, 2016; Ruffle and Wilson, 2018, 2019). However, tattooed individuals are still stigmatized and face discrimination (Jibuti 2018, Dolaec and Stein, 2013). By examining the impact of visible tattoos on employment chances in a sector characterized by high skill requirements but low in-person customer contact[44], this paper further extends the literature examining labor market discrimination against controllable characteristics.

[43] There may also be other reasons than the lack of power behind insignificant results. However, power calculations suggested for the effect size about 10 percentage points, approximately 800 observations were needed.
[44] Therefore, customer discrimination concerns should be minimal in this setting.

I used a correspondence testing method to collect data and responded to 800 job advertisements in IT occupations in Germany with fictitious applications. Applications included a large size photo on the cover page with or without a neck tattoo, to signal the tattoo status of applicants. Our data suggest that applicants with visible tattoos have about a 25% lower chance of getting a callback compared to applicants without tattoos. The negative impact of a visible tattoo remains significant even after controlling for company/position related characteristics and monthly and regional fixed effects. I tested whether applicants face statistical discrimination but failed to find evidence for it. While examining within group variation of the effect, I found that discrimination against tattooed applicants was largely similar across those groups (male/female, eastern/western Germany, urban/rural area and for different firm sizes), since none of the interaction terms are statistically significant. However, the treatment effect becomes insignificant in more than half of the regression specifications; thus, my analyses lack statistical power to argue about the similarities of tattoo impact across various subgroups.

Since no evidence was found to confirm that tattooed applicants suffer from statistical discrimination, I argue that applicants with visible tattoos face taste-based discrimination. Alternatively, HR managers may exhibit hiring bias against tattooed candidates, based on their beliefs about personal characteristics of individuals with tattoos, which may be motivated by animus. This result is consistent with Jibuti (2018), who also argues that tattooed applicants face taste-based discrimination in the banking industry in Germany. Despite the growing acceptance of tattoos among the wider public (Walzer and Sanjurjo, 2015), employers in the banking and IT sector still see it as a hindrance to employment. The fact that a high level of discrimination is observed in both the banking and IT sector supports our argument that the discrimination is a result of a distaste for tattoos. One would expect unfavorable treatment of tattooed applicants in the front-office jobs in the banking industry, as appearance is an important factor because of the extent of interaction of workers with customers. However, the IT sector is more skill-intensive with minimal face-to-face contact with clients. Hence, if tattoos negatively affect the appearance of workers, they should not affect callback rates in the IT sector. Moreover, increasing the skill level did not reduce or eliminate discrimination in the IT industry. Therefore, even if the wider public appears to have become more tolerant of tattooed individuals, employers still seem to be less receptive to tattoos in the workplace.

I should highlight that our results are context specific. I do not argue that all types of tattoos would have a similar impact on employment chances in all occupations. Rather, my results are conditional on neutral tribal tattoos in the IT industry. On the other hand, given that Jibuti (2018) found the unfavorable treatment of tattooed applicants in the banking sector, one may argue that tattoos may still have a rather homogeneous impact across diverse settings. Still, if we want to fully understand the impact of visible tattoos on employment chances, a broader range of occupations should be examined and the content or location of tattoos varied. Some occupations, such as art-related professions (actor, designer) may reward tattooed workers, as tattoos may ascertain artistic affiliation, translating into a favorable treatment. Moreover, tattoos with specific connotations (army, religion) may trigger different attitudes from HR managers, compared to neutral tribal tattoos, leading to a different effect. Finally, HR managers' implicit attitudes on acceptance of tattoos in the workplace could potentially be captured using the Implicit Association Test (Greenwald et al., 1998). Then the link between the IAT scores and actual behavior could be examined to find whether implicit attitudes are behind hiring bias. Therefore, further research in this direction could answer questions of whether tattoos are uniformly unaccepted in any occupation and what the driving force is behind HR managers' distaste for tattoos.

3. Race Gender and Beauty: The Effect of Information Provision on Online Hiring Biases[45]

co-authored with Weiwen Leung[46] Zheng Zhang[47] Daviti Jibuti[48] Jinhao Zhao[49] Maximillian Klein[50] Casey Pierce [51] Lionel Robert [52] Haiyi Zhu[53]

3.1 Introduction

This paper examines the prevalence of different types of biases in hiring. To collect data, we conducted a study by setting up a task on Amazon MTurk in which we recruited 206 subjects. These subjects made hiring decisions in a platform simulating a website recruiting people for freelance jobs. We examine first and foremost how hiring rates are affected by gender, race, and beauty. Next, we explore whether the number of people displayed and/or performance information affects hiring decisions.

Labor market discrimination based on race and gender has been widely examined in discrimination literature (Neumark, 2018). Some studies have also examined the impact of beauty on employment chances (Katuscak and Kraft, 2013; Ruffle and Shtudiner, 2015). However, there is little known about how increasing the size of the choice set impacts hiring decisions. Theoretical and empirical studies have shown that increasing the size of the choice set reduces agents' participation in the market (Iyengar and Kamenica, 2010), particularly if the available choice set contains useful information (Kamenica, 2008). Due to this *choice overload* phenomenon, agents prefer simpler options and are more likely to use heuristics in decision-making (Kahneman, 2011). These cognitive biases usually lead to formation of implicit preferences, which can predict behavior including hiring discrimination (Rooth, 2010). Therefore, increasing the size of the choice set can result in discrimination in hiring through the process described

---

[45] We are grateful to Christine Exley, Max Harper, Loren Terveen, Brent Hecht, Teng Ye, and Jacob Thebault-Spieker for their help at different stages of the project. This work was supported by the National Science Foundation, under grant IIS-2001851, grant IIS-2000782, and grant IIS-1939606. ACM ISBN 978-1-45036708-0/20/04 ...$15.00. DOI: https://dx.doi.org/10.1145/3313831.3376874.

46 Email: iamweiwenleung@gmail.com
47 University of Rochester. Email: zzhang95@cs.rochester.edu
48 CERGE-EI, a joint workplace of Charles University and the Economics Institute of the Czech Academy of Sciences, Politickych Veznu 7, 111 21 Prague, Czech Republic. Email: Daviti.Jibuti@cergeei.cz
49 Email: jinha14@tsinghua.org.cn
50 Email: max@notconfusing.com
51 University of Michigan. Email: cbspierc@umich.edu
52 University of Michigan. Email: lprobert@umich.edu
53 Carnegie Mellon University. Email: haiyiz@cs.cmu.edu

above. To test whether increasing the number of available candidates has an impact on hiring preferences, we vary the number of potential applicants in one of our treatment manipulations and observe whether this alters hiring managers' decisions.

Another experimental manipulation, via the provision of additional information, is usually used to disentangle theories of statistical or taste-based discrimination (Bertrand and Duflo, 2017). When additional productivity-related information eliminates or reduces bias against disadvantaged groups, researchers usually interpret this as evidence of statistical discrimination (Phelps 1972). If bias is unaffected by additional information, taste-based discrimination (Becker 1957) may be prevalent. Therefore, providing additional performance information would inform us of the underlying mechanisms that cause discrimination. Guryan et al. (2013) note that this is an important question in the economics literature on discrimination, which has existed for over 50 years.

Racial discrimination in offline hiring has been well documented, especially in the US. In particular, there is significant discrimination against African Americans and Latinos in hiring (Bertrand and Mullainathan, 2004; Quillian et al., 2017). Biases based on gender (Bohnet et al., 2016) and beauty (Mobius and Rosenblat, 2006) are also prevalent. However, there are good reasons to suspect that online platforms may lessen or eliminate hiring biases. For example, Morton et al. (2003) use observational data to show that while racial minorities pay 2% more for cars when purchasing them offline, this gap is much smaller for online purchases. The authors attribute this to the internet facilitating information search and removing cues present in offline negotiations. To the extent that such considerations are applicable, the internet may have a similar impact on the digital gig market.

Studying hiring bias in online platforms is also of policy relevance. An increasing number of Americans are earning money through freelance jobs obtained via online platforms. Indeed, a report from the Pew Research Center (Smith 2016) indicates that 8% of Americans earn money from these "digital gigs". Websites that host these services facilitate supplemental income for some workers and become a primary income source for others. The report also states that 14% of Black respondents and 11% of Latino respondents reported earning money on these platforms during the previous year, in contrast to 5% of White respondents. Among these non-White workers, 65% of them describe the income they earn from these platforms as “essential” or “important”. Additionally, 55% of online gig workers are female. One important question is the extent to which different types of hiring biases exist on these platforms, both with respect to easily quantifiable characteristics such as race and gender, but also less easily quantifiable characteristics such as beauty.

Our experiment results suggest hiring biases against Black candidates, and towards Asian and female candidates. Specifically, Black candidates are hired 16% less than White candidates, while Asian candidates are hired 23% more than White candidates. Regarding gender, females are hired 61% more than males and there is no significant difference between White and Latino candidates. In relation to beauty, we find hiring biases toward more attractive candidates. A one standard deviation increase in attractiveness increased hiring chances by around 10%.

The next experimental manipulation includes showing the prior performance of candidates in a similar task. We expected this additional information to reduce or eliminate discrimination. Our data suggests that the provision of information at the individual level erased the differences between White, Asian and Black candidates, and the difference between attractive candidates versus non-attractive candidates. On the other hand, additional information further increased the hiring chances of female candidates, compared to males. When the information was provided at the subgroup level by gender, the hiring chances of female candidates did not decrease. However, it reduced the hiring chance for Black candidates and increased the hiring chance for Asian candidates.

Lastly, we also manipulated the number of candidates our employer subjects could choose from, to test how the choice overload impacts hiring decisions. As we conjectured, increasing the number of candidates to choose from reduced the chance that Black candidates would be hired. Asians are hired more frequently when the number of available candidates increases from 2 to 4 and the difference is statistically significant. As for Latino candidates, their hiring chance reduces as the number of candidates increases; however, the effect is not significant. Additionally, we found that female applicants are hired significantly more often when the number of available candidates increases, while the impact of attractiveness is broadly stable. We subsequently discuss our interpretation of the results.

The paper is organized as follows. The next section will review more closely the existing literature on discrimination, and provides a base for our development of hypotheses for testing. In the subsequent sections, we describe the experimental design and data analysis methodology. Section 5 describes our findings, while section 6 provides potential interpretation and practical implications. Section 7 concludes.

3.2 Literature Review and Hypotheses

3.2.1 Bias and Discrimination in Hiring

Since the seminal work by Bertrand and Mullainathan (2004), the discrimination literature has seen strong growth. Key areas of study have been racial and gender discrimination. Edelman et al. (2017) find that people with distinctively Black names are 16 percent less likely to be accepted as guests on AirBnB compared to those with distinctively White names. Pope and Sydnor (2011) study online loans on the peer-to-peer website Prosper.com, and find that loan listings with Blacks[54] in the attached picture are 25 to 35 percent less likely to receive funding than those of Whites with similar credit profiles. However, despite the higher average interest rates charged to Blacks, lenders making such loans earn a lower net return than for loans made to Whites with similar credit profiles because Blacks have higher relative default rates. Other evidence consistent with racial discrimination in online environments includes an observational study by Hannak et al. (2017) who found that Black people tend to get more negative reviews than other races, which could harm their employment opportunities. Bartos et al (2016) conducted a series of field experiments and found that Asian minorities are discriminated against compared to white majorities in the labor market and the rental housing market. Moreover, employers put more effort into accessing resumes of white majority candidates, while landlords acquire more information about Asian minority applicants. The authors claim that this is consistent with attention discrimination theory, where the selectivity of the market affects the optimal level of attention devoted to different groups.

Gender discrimination has also been studied, and a resulting theme suggests that whether or not females are discriminated against depends heavily on the task and context. For example, Bohnet et al. (2016) find pro-female discrimination in hiring on language

[54] We use "Black" instead of "African-American" to be consistent with the original study. In the rest of the paper, we use terminology that is consistent with the underlying sources as far as possible.

tasks and anti-female discrimination in mathematics tasks when candidates are evaluated one at a time. However, discrimination disappears when candidates are evaluated jointly. Coffman et al. (2019) study gender discrimination when candidates are evaluated two at a time for sterotypically male tasks, and find discrimination when two candidates' prior performances are equal, but not when there is a candidate with a stronger prior performance. Finally, a field experiment on mathematics Stackexchange (Bohren et al., 2019) found that low-reputation users with female usernames receive fewer upvotes for questions they post relative to those with male usernames. However, the direction of discrimination reverses at high reputation levels: those with female usernames receive more upvotes. The authors explain their findings could be due to people having incorrect beliefs about female math ability. Interestingly, there is no evidence for gender discrimination with regards to posted answers, and the authors attribute it to the decreased subjectivity over whether answers should be upvoted (as compared to questions). Gender discrimination can also vary over time; Tang et al.'s (2017) study of LinkedIn data found that gender discrimination has decreased significantly over the past 10 years.

After the seminal work by Hammermesh and Biddle (1994) vast body of literature examines how decision-makers are affected by attractiveness. In a heavily cited lab experiment, Mobius and Rosenblat (2006) find a sizable beauty premium in hiring, as physically attractive workers are more confident and considered more able by employers, and are also thought to have better oral skills. Jenq et al. (2015) study an online charitable microfinance website and find that borrowers who are more attractive receive funding more quickly. In a correspondence study, Katuscak and Kraft (2013) find that more attractive applicants enjoy higher callback rates and there was a slightly higher beauty premium for male applicants.

In this paper, we explore the extent to which different forms of hiring biases based on gender, race, and attractiveness can manifest themselves in an online freelancer marketplace, and then examine the effect of information provision on hiring biases. We focus on math as our task domain because race-based and gender-based stereotypes on math are well-documented in the literature (Ellemets, 2018). Furthermore, multiple sources indicate gender and racial gaps in SAT math scores that have persisted over time.

In particular, males outperform females[55], Asians outperform Whites, and Whites outperform both Blacks and Latinos[56].

Based on these, we formulate our hypotheses H1 to H5, all of which are conditional on observable characteristics.

H1. *Females will be hired less frequently than males*.

H2. *Asians will be hired more frequently than Whites*.

H3. *Whites will be hired more frequently than Blacks*.

H4. *Whites will be hired more frequently than Latinos*.

H5. *A more beautiful person will be hired more often than a less beautiful person*.[57]

As described above, hypotheses H1 and H2 are related to stereotypes, and hence this bias would be consistent with the theory of statistical discrimination. Similarly, although in general Blacks and Latinos may face taste-based discrimination, in the context of our experiment settings, H3 and H4 would also be linked to statistical discrimination. On the other hand, hiring bias in H5 can be related to either statistical or taste-based discrimination, as the related literature does not clearly link the behavior to a particular theory.

3.2.2 Information Provision

There are many ways in which provision of additional information can affect behavior. For example, certain subgroups may be less likely to be hired as they are perceived to be less productive than other subgroups, a phenomenon known as statistical discrimination (Phelps, 1972). However, the provision of information on individuals' performance on previous tasks can reduce statistical discrimination (Bertrand and Duflo, 2017).

[55] http://www.aei.org/publication/2016-sat-test-results-confirm-pattern-thats-persisted-for-45-years-high-school-boys-are-better-at-math-than-girls/, https://www.fairtest.org/sat-act-gender-gaps

[56] https://www.brookings.edu/research/race-gaps-in-sat-scores-highlight-inequality-and-hinder-upward-mobility/, https://www.insidehighered.com/news/2017/09/27/scores-new-sat-show-large-gaps-race-and-ethnicity

[57] We are not aware of any data that examines the correlation between beauty and math test scores. However, the H5 is based on the studies which found more favorable outcomes for beautiful candidates (Jenq et al., 2015; Mobius and Rosenblat, 2006; Katuscak and Kraft, 2012).

Similarly, the effect of information on how certain subgroups performed can also help disadvantaged groups. For example, if subgroup A is hired more often than subgroup B, but information reveals that both subgroups are equally productive, then information about subgroup performance should reduce the gap. On the other hand, if both subgroups are hired equally often, but information reveals that subgroup A is more productive, then subgroup performance information should result in workers from subgroup A being hired more often. In our experiment, we examine how the provision of information at the individual level (how the candidate did in previous tasks) and subgroup level (how one's subgroup did in previous tasks) affect hiring decisions and formulate the next set of hypotheses, H6 and H7, which can be linked to the theory of statistical discrimination.

H6. *Provision of information at the individual level (how the candidate did in previous tasks) can reduce hiring bias*.

H7. *Provision of information at the subgroup level (how the candidate's subgroup did in previous tasks) can reduce hiring bias against disadvantaged groups*.

Another way that information can affect decision making is by altering the choice environment by using behavioral ``nudges'' (Thaler and Sustein, 2009). For example, Lee et al. (2011) find that behavioral economics persuasion techniques, such as having default options, can lead to people making healthier food choices. One well-known nudge is to vary the number of options to choose from (i.e. the size of the choice set). Iyengar and Lepper (2000) showed that people are much more likely to buy jam when faced with 6 varieties than when faced with 24 varieties, a phenomenon known as ``choice overload''. While we know that increasing the number of options makes one less likely to make a choice (Cherneva et al., 2015), what is less well known is the effect on which choice is made. Our study addresses this issue by providing more insight into how choice overload affects which choice is made, with a focus on equity concerns. To the best of our knowledge, our paper is the first work that studies the impact of choice size on discrimination.

In the context of online hiring, we propose that the size of the choice set can influence hiring biases. One natural hypothesis may be that increasing the number of candidates for hire may lead people to use heuristics - gender- or race-based stereotypes. Indeed, under Kahneman's dual-system framework, people are more likely to use heuristics when overloaded with information (Kahneman, 2011). One of the few studies examining the impact of choice set size, found that people tended to go with easy-to-understand (e.g. less risky) options when the choice set expanded (Iyengar and Kamenica, 2010). We hypothesize that people are more likely to use heuristics that will accentuate existing biases (e.g. those based on stereotypes) when faced with a larger choice set. These "automatic preferences" can be captured by the Implicit Association Test (IAT) developed by Greenwald et al. (1998) and there is some evidence that implicit attitudes can predict actual discriminatory behavior (Rooth, 2010). Based on these considerations we formulate our eighth hypothesis as below.

H8. *Increasing the number of candidates to choose from can increase hiring bias*.

3.3 Experimental Design

We designed an experiment where participants were told they would be making hiring decisions for a mathematically intensive task. We informed participants[58] that potential employees had completed two sets of mathematical tasks, one easy set and one difficult set (Round 1 and Round 2, respectively). Both sets had five questions each. We showed employer subjects example questions from both sets. The easy questions were similar in difficulty to easy SAT questions, and the difficult questions were similar in difficulty to difficult SAT questions.

Our experiment was designed around mathematically intensive tasks for several reasons. First, clear stereotypes exist, at least with regards to gender (Ellemers, 2018). Second, the discrimination literature often uses mathematically intensive tasks as a subject of study (Bohnet et al., 2016, Bohren et al., 2019, Coffman et al., 2019). Finally, many gig work tasks involve the use of mathematics: a search of sites such as Fiverr and Upwork reveal thousands of math-related tasks.

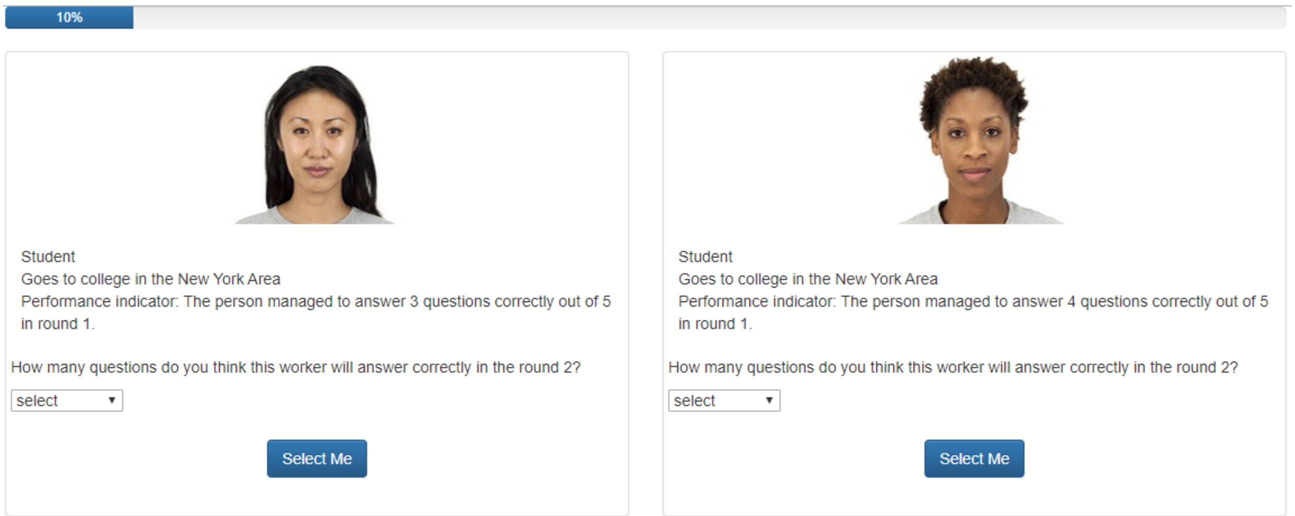
All participants were told there would be twelve hiring rounds, and they would make one hiring decision in each round with the photos of potential employees also displayed. A third of the participants were informed that they would see two potential employees in each round, while another third would see four, and the remaining third would see eight. Figure 3.1 shows a screenshot of the condition with two potential employees. As shown in the figure, we also asked participants to predict the number of difficult questions each worker would answer correctly, before each hiring decision. This

[58] We recruited U.S. MTurkers who had completed at least 500 tasks and an acceptance rate of at least 97%

technique is known in the discrimination literature as “belief elicitation“ (see e.g. Coffman et al. (2019)) and is used to examine whether discrimination (if present) is due to people's beliefs about the productivity of different subgroups.

Before the hiring experiment, we employed 106 subjects on MTurk, who performed 2 rounds of a worker's task (a mathematically intensive task). For signaling gender, race and attractiveness we decided to use photographs. However, we could not use photos of those subjects who actually solved mathematical tasks because MTurk does not allow us to take or request photos from MTurkers. For this reason, we decided to use photos from the Chicago Face Database, so that we could obtain measures of perceptions of race, gender, and attractiveness. We ensure that the gender of the photo corresponded to the gender of the potential employee we hired from MTurk to answer SAT-level math questions.

Figure 3.1: Screenshot of two-worker condition with performance on easy questions displayed.

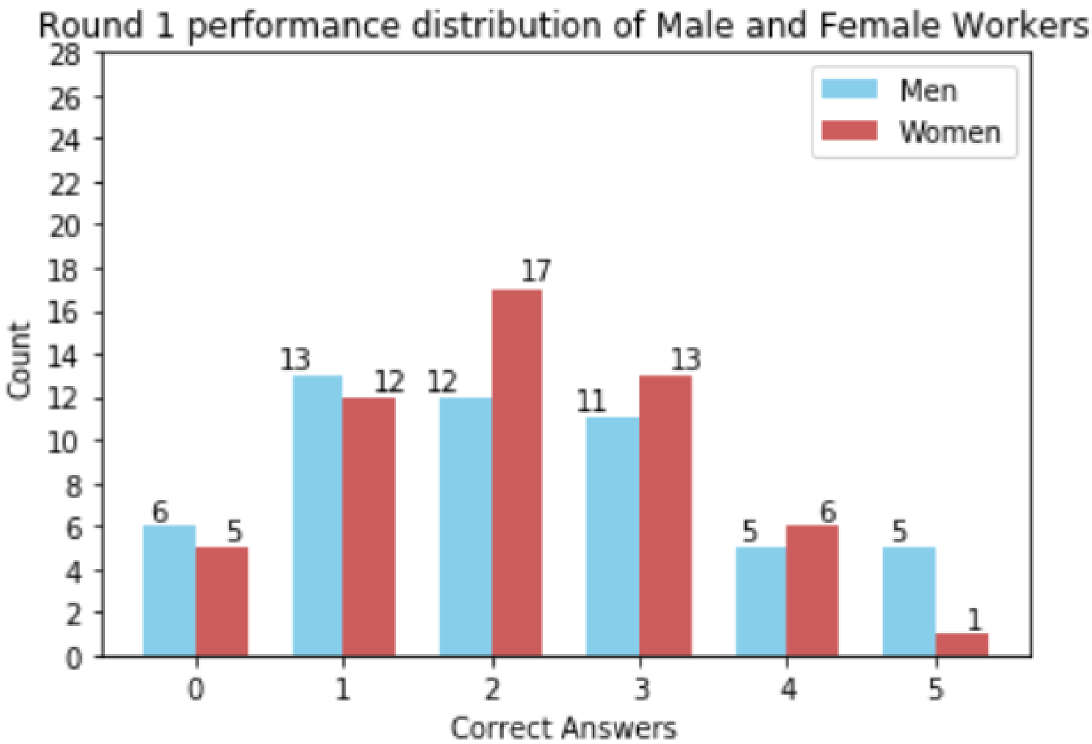


Participants were incentivized to take hiring decisions seriously: they were told that after they made all hiring decisions, one of their hires would be randomly selected, and they would be given a bonus of $1 for every question the person they hired on a randomly selected round had correctly solved on the difficult set[59]. The average payment to the participant was $2.8 which included base pay of $1 and an average bonus of $1.8.

[59] There was a total of five questions, so the maximum bonus payout was $5 (in addition to the participation payment).

A portion of subjects was provided with additional information before they made any hiring decisions. Specifically, a third of all participants saw the number of questions that potential employees correctly solved on the easy set (Figure 1 shows a screenshot of a trial with such information). Another third saw the performance distribution by gender of questions that potential employees correctly solved on the easy set in Round 1 (screenshot available in Figure 3.). The remaining third of participants did not see either.

Figure 3.2: Distribution of performance by gender, which was shown to randomly selected participants before they made their hiring decisions, along with a short explanation of how to interpret the distributions presented.



Because the photos shown to participants were chosen randomly from the Chicago Face Database (orthogonal to other characteristics, except with respect to gender), the expected past performance of potential employees of each race was equal. Likewise, the expected past performance of potential employees was unaffected by their beauty. In addition, we matched potential workers in the experimental trials so that the average past performance (displayed to participants) across genders would be exactly equal[60]. In this way, calculating the average number of times a worker from a subgroup was chosen (without controlling for covariates) and comparing that to other subgroups gives an

[60] Appendix C.4 provides more details on experimental trials and shows how we ensure that past performance by gender in each condition is equal.

estimate of hiring preferences. Therefore, if the race, gender, and beauty of a potential employee were irrelevant to our participants, we should discover that these factors had no effect on hiring.

All participants were given comprehension questions to make sure they understood the nature of the experiment (including that their payout would depend on the hiring decisions they made) and had to answer the questions correctly before they could proceed with hiring decisions.

3.4 Data Analysis Methodology

We use discrete choice modeling to analyze our data. A discrete choice model is a form of agent-based model that is often used in economics (Nevo, 2000), marketing (Ben-Akiva and Boccara, 1995), transportation (Ben-Akiva and Bierlaire, 1999), and public health (Lancsar and Louviere, 2008), among other fields. In a discrete choice model, a decision-maker is assumed to choose between different alternatives (e.g. products, healthcare options, transportation options, or in our case, potential employees) by computing the value of each alternative as a function of that alternative's characteristics.

Suppose that a decision-maker is considering N alternatives (in our case, potential employees). In our case, a decision-maker might compute the value of potential employee i ($i \in \{1,2,...,N\}$) follows:

$$Value_i = \beta_0 + \beta_1 * Female_i + \beta_2 * Asian_i + \beta_3 * Black_i + \beta_4 * Latino_i + \beta_5 * Attractiveness_i$$

where $Female_i$, $Asian_i$, $Black_i$, $Latino_i$ are variables indicating the gender and race of the worker. $Attractiveness_i$ is a continuous variable measuring the attractiveness of the worker. We did not ask our participants to evaluate the attractiveness, race, or gender of each potential employee (and doing so would be time consuming and interfere with participants' decisions[61]). Instead, we proxied these variables by using their values from the corresponding photo in the Chicago Face Database, which was based on the results of a survey on the proportion of people who thought the person in the photo was female, Asian, Black[62], or Latino, as well as the average attractiveness rating of the photo. This

61 However, we did ask participants to predict the number of difficult questions each participant would answer correctly, because such information was valuable and could not be proxied by data from elsewhere.
62 To be consistent with Chicago Face Database terminology, we use "Black" and not "African-American"

introduces measurement error, but classical measurement error biases our coefficient estimates towards zero, making it harder for us to find effects that in fact exist (Wooldridge, 2019).

Decision-makers want to choose the option with the highest value. However, they measure value with error e.g. because of errors in perception, errors in computation, or due to randomness in taste. The chance that they will choose a particular option is therefore a probabilistic function that increases as the value of that particular option increases and decreases as the value of alternative options increases. The exact mathematical equations governing our model can be found in the Appendix.

The discrete choice model has several desirable properties. Perhaps most importantly, it allows for decision makers to take into account the *relative* value of each alternative when making a decision. For example, an option with a value of 10 would likely be chosen if there was only one alternative option with a value of 1, but not if the alternative option had a value of 100[63]. The model also flexibly adjusts to the fact that the probability of choosing any alternative decreases when more choices are available, which is important for our case since we vary the number of potential employees.

We use a conditional logit model (McFadden, 1973) and a maximum likelihood estimation (MLE) to estimate the parameters (β's) in the model. We cluster standard errors by participant because each participant makes multiple decisions. Estimating the model using our entire sample allows us to estimate the overall effects of race, gender, and attractiveness. To evaluate the effects of these variables under each experimental manipulation, we estimate the model using only data from participants who were exposed to that experimental manipulation.

In the discrete choice model, our decision makers estimate value. Thus our coefficient estimates should be interpreted as the marginal value of a given attribute. The effect on the probability of hiring can be computed through the use of odds ratios[64]. Note that since the probability of hiring is monotonically increasing in value, positive

[63] A standard linear/logistic regression only makes use of a given alternative's characteristics, and it would be impossible or extremely difficult to replicate the flexibility of the discrete choice model by adding control variables, especially since the size of the choice set varies across participants.
[64] The odds ratio of a coefficient estimates of X is e^X. Specifically, if a coefficient estimate of X indicates that a one unit increase in the explanatory variable is associated with a e^X-1 increase in the probability of hire, holding other explanatory variables constant. For example, $X = 0.1$ would correspond to a roughly *10.5*% increase.

coefficient estimates always indicate a positive effect of a given attribute on the probability of hire.

Our experiment was designed such that the workers from the different subgroups of interest were in fact on average equally qualified in terms of observable characteristics (prior performance). Hence, we can make claims regarding equally qualified workers (in terms of those observable characteristics) without controlling for previous performance. We do not control for previous performance because we do not display the previous performance of workers to half of our participants.

Before we discuss the results, we briefly note two limitations of our methodology. First, even though attractiveness was measured by independent coders engaged by the Chicago Face Database, notions of attractiveness may reflect Western concepts. Second, while our data covers male and female genders well, we may not be able to generalize to other genders.

3.5 Results

3.5.1 Descriptive statistics

Table 3.1 shows various descriptive statistics associated with potential employees. The first row in Panel A indicates performance in Round 1 (easy SAT math questions) by gender of the candidates that we hired to answer math questions. We see that males perform better in Round 1 than females; however, the difference is insignificant. As the table shows, females answered, on average, 2.04 questions correctly, while males managed to answer 2.21. Overall in Round 1 workers had 5 questions to answer.

The second row indicates the mean predicted score of candidates[65] given by participants in our experiment. Participants expected females to perform significantly better than males in Round 2 (harder SAT math questions). One possible explanation behind the higher prediction for female workers is that participants may have thought that gender discrimination was the topic of study and tried to counteract any implicit biases they held, despite being incentivized. This explanation is supported by a post-

[65] Candidates that appeared to participants in our experiment. While we did hire MTurkers to solve easy and difficult math questions, we were not allowed to take photos of them, or ask them to supply photos. So we took photos of people from the Chicago Face Database, and matched them with the workers we hired based on gender information.

experimental survey (discussed below), in which the majority of subjects associate math with males. Note also that predicted scores could have been higher for harder SAT math questions than actual scores for easy SAT questions because most participants did not see the distribution of scores for easy SAT questions.

Table 3.1: Descriptive statistics of potential worker

Panel A – Gender	Female	Male	*p-value*	
Round 1 performance	2.04 (0.16)	2.21 (0.20)	0.49	
Prediction	2.74 (0.02)	2.65 (0.02)	0.001	
Panel B – Race	Asian	Black	Latino	White
Prediction	2.80 (0.04)	2.60 (0.04)	2.76 (0.04)	2.69 (0.02)
*p-value**	0.005	0.007	0.077	
Panel C - Attractiveness	Female	Male	p-value	
Attractiveness score	3.44 (0.01)	3.08 (0.01)	0.00	

Notes: Standard errors in parenthesis beneath mean estimates. The last column shows p-values of the hypothesis of equal means across groups. * - compares means of relevant race groups to mean of *Whites*

Panel B shows how predicted scores differ across race: *Asians* had the highest expected Round 2 scores (2.8), followed by *Latinos*, *Whites* and *Blacks* (2.76, 2.69 and 2.6, respectively). When comparing the predicted performance of each group to the base group (i.e., *Whites*), we see that the difference is highly significant for *Asians* and *Blacks* (p-value less than 0.01), while it is only marginally significant for *Latinos* (second row in Panel B). This is broadly in line with the findings with regards to race and SAT scores, described above, where *Asians* outperform *Whites* and *Whites* outperform *Blacks*.

Panel C of Table 3.1 shows the mean attractiveness score of photos that appeared in our experiment (by gender) as given by coders in the Chicago Face Database. Females are perceived to be significantly more attractive than males. Experiment subjects expected more attractive candidates to have on average 0.13 higher scores (p-value = 0.00) in

Round 2 than less attractive applicants[66]. This is also consistent with our expectation that more attractive candidates may be viewed as more employable than less attractive ones.

We now focus on participants who made hiring decisions (i.e., employers). They came from a wide variety of backgrounds. For example, they live in 34 different U.S. states; the three states which contributed the most number of participants had 18%, 12%, and 8% of the subject pool, and all other states each contributed 4% or less. 42% of the participants are females and the rest are males. In terms of their highest educational level, 3% of our sample have a high school diploma or lower and 17% have either some college or a 2-year college degree. 45% of our subjects have 4 years of college degree, 29% have a master's degree and 6% have a professional degree. Although the sample skews towards the more educated, one might expect that the more educated are more likely to hire people in the gig economy due to higher income. Unfortunately, we did not collect data on race or mathematical ability.[67]

When asked explicitly whether math was associated with a particular gender, subjects tended to associate math with males (rather than females). Only 17% of our sample somewhat associate math with females, with 27% not associating math with any gender and the remaining 56% associating it with males. On the other hand, approximately equal numbers of subjects associate liberal arts with females (35%) and males (32%), while a third of participants (33%) think they are not related to any specific gender. The demographics and opinions were collected at the end of the study, after subjects completed the hiring rounds.

In terms of race of the people in the photos we took from the Chicago Face Database, almost a third of workers were White (31%) followed by Blacks (28%), Asian (22%) and Hispanics (19%).[68]

3.5.2 Discrete Choice Model: Results with Full Sample

Estimating on the full sample indicates that attractive candidates are valued more and hence hired more often than equally qualified unattractive candidates, as evidenced by the positive and statistically significant coefficient of *Attractive_i*. This confirms our

[66] This result is not shown in Table 1.
[67] It might be useful for future work to examine the influence of mathematical ability on discrimination.
[68] We define race in line with the Chicago Face Database definition.

hypothesis of a beauty premium. Similarly, Whites are hired more often than Blacks and Latinos, while Asian candidates are hired more frequently than White candidates (male and White are omitted categories; we do not mention the omitted categories from now on for brevity). These results are in line with our hypothesis of preferential treatment of Whites over Blacks/Latinos and Asians over Whites. Contrary to our expectations, however, females have a significantly higher probability of being hired, relative to male candidates (see leftmost column of Table 3.2).

3.5.3 Showing Prior Performance

When we do not show any performance information on easy mathematics questions, all explanatory variables that were significant in the full sample remain significant and have the same sign, except for Black_i. Results in a column with the heading "None" indicate that Black candidates are not chosen at a different rate than White candidates. Since this column shows results without any performance information, we interpret it as prior preferences of our employer subjects.

The next column in Table 3.2 shows results when we show candidates' individual performance on easy mathematics questions. The results indicate that attractiveness and race no longer has a statistically significant effect on hiring. Female candidates are still hired more often, and in fact the coefficient estimate of Female increases compared to when no information about prior performance is displayed. These results indicate that preferences towards attractive and Asian workers can be explained by statistical discrimination (Phelps, 1972).

Evidence of statistical discrimination is confirmed in the rightmost column of the table, which extends the regression model in the second column to include interactions of various worker characteristics with treatment dummies. Coefficient estimates of interaction of individual information treatment with the variables *Attractiveness* and *Asian* are negative and statistically significant. This suggests that once information on the individual level performance is provided, (positive) discrimination against attractive and Asian candidates is eliminated, consistent with the theory of statistical discrimination. On the other hand, preferences towards females may be driven by taste-based discrimination (Becker, 1971), as providing individual-level information further increases hiring chances for females. The last column reaffirms this result as evidenced by a positive and significant interaction term. As for Blacks and Latinos, this treatment does not have a significant impact on the hiring chances of these groups. Therefore, we conclude that our hypothesis about provision of additional individual-level information potentially reducing bias is largely or completely supported, depending on how one interprets the further increase in likelihood of female candidates being hired.

Table 3.2: Estimates of Discrete Choice Model – Information provision treatment

Dependent variable: *Choice*	Full sample	Prior performance shown			
		none	*individual*	*subgroup*	*All*
Attractiveness	0.14*** (0.03)	0.24*** (0.06)	0.03 (0.06)	0.16*** (0.06)	0.24*** (0.06)
Attractiveness * Individual info treatment					-0.21*** (0.08)
Attractiveness * Subgroup info treatment					-0.08 (0.08)
Female prop	0.48*** (0.06)	0.36*** (0.09)	0.67*** (0.10)	0.42*** (0.10)	0.36*** (0.09)
Female prop * Individual info treatment					0.31** (0.13)
Female prop * Subgroup info treatment					0.06 (0.14)
Asian prop	0.21** (0.08)	0.30** (0.15)	-0.07 (0.15)	0.36** (0.14)	0.30** (0.15)
Asian prop * Individual info treatment					-0.37* (0.21)
Asian prop * Subgroup info treatment					0.07 (0.21)
Black prop	-0.17** (0.07)	0.03 (0.12)	-0.18 (0.12)	-0.38*** (0.13)	0.03 (0.12)
Black prop * Individual info treatment					-0.22 (0.18)
Black prop * Subgroup info treatment					-0.41** (0.18)
Latino prop	-0.12 (0.12)	0.20 (0.20)	-0.21 (0.22)	-0.37 (0.23)	0.20 (0.20)
Latino prop * Individual info treatment					-0.42 (0.30)
Latino prop * Subgroup info treatment					-0.58* (0.30)
N	9256	3320	2846	3090	9256
Number of clusters	2216	796	703	717	2216
Pseudo R^2	0.025	0.026	0.032	0.031	0.030

Notes: Standard errors are clustered at the subject level. * p<0.1, ** p<0.05, *** p<0.01.

The fourth column in Table 3.2 shows results in a group that prior to hiring rounds saw the distribution of candidates' performance on easy questions by gender. Even though women actually performed slightly worse than men on the easy SAT-level math questions, compared to no information on prior performance (column“ *None*“), the chances of female candidates being hired increases slightly in this treatment. However, the difference is not statistically significant at conventional levels, as the coefficient of the relevant interaction term suggests in the last column. Regarding race, the information on subgroup performance by gender significantly increased hiring bias against Black and Latino candidates. The results in the last column show that interaction terms with subgroup information treatment are negative and statistically significant for these groups, indicating lower hiring chances. These results imply that, contrary to our expectations, provision of information at the subgroup level did not reduce hiring bias. Interestingly, the hiring chances increase for Asian candidates and decrease for attractive candidates in this treatment manipulation, though the treatment effect is insignificant in both cases (see the relevant interaction coefficients in the last column).[69]

3.5.4 Number of Candidates

Among participants who were asked to choose between two candidates at a time, we find that only attractive candidates and females are hired significantly more often. The variables indicating race are not statistically significant (see Table 3.3, column "2").

In the treatment where participants choose between four candidates at a time, they chose attractive, female and Asian candidates significantly more often. However, as the last column shows, the treatment effect is insignificant for attractive and Asian candidates. On the other hand, compared to the two-worker condition, females are hired significantly less often, the only treatment effect which is statistically significant. Black candidates are chosen less often than Whites, but the difference is not statistically distinguishable from zero (column "4" in Table 3.3).

[69] Analyses of sections 5.3 to 5.6 were also performed to disaggregate by the gender of employer subjects. We find similar patterns in decisions of female and male recruiters, with effects somewhat stronger for male recruiters. The only noticeable difference was for *Attractiveness*, with male recruiters expressing strong preferences towards attractive candidates, while attractiveness did not have a significant impact on decisions of female recruiters. Detailed tables can be found in Appendix E.

Table 3.3: Estimates of Discrete Choice Model – Number of available applicants' treatment

Dependent variable: *Choice*	Number of displayed candidates			Full sample
	2	*4*	*8*	
Attractiveness	0.14*** (0.06)	0.13*** (0.06)	0.15*** (0.06)	0.14** (0.06)
Attractiveness * 4 worker treatment				-0.01 (0.08)
Attractiveness * 8 worker treatment				0.01 (0.08)
Female prop	0.55*** (0.10)	0.30*** (0.09)	0.62*** (0.10)	0.55*** (0.10)
Female prop * 4 worker treatment				-0.26* (0.14)
Female prop * 8 worker treatment				0.07 (0.14)
Asian prop	0.10 (0.15)	0.35** (0.14)	0.15 (0.14)	0.10 (0.15)
Asian prop * 4 worker treatment				0.25 (0.21)
Asian prop * 8 worker treatment				0.05 (0.21)
Black prop	-0.08 (0.13)	-0.13 (0.12)	-0.29** (0.13)	-0.08 (0.13)
Black prop * 4 worker treatment				-0.05 (0.18)
Black prop * 8 worker treatment				-0.21 (0.18)
Latino prop	-0.07 (0.23)	0.02 (0.20)	-0.23 (0.22)	-0.07 (0.23)
Latino prop * 4 worker treatment				0.06 (0.30)
Latino prop * 8 worker treatment				-0.16 (0.31)
N	1912	2736	4608	9256
Number of clusters	956	684	576	2216
Pseudo R^2	0.035	0.017	0.031	0.027

Notes: Standard errors are clustered at the subject level. * p<0.1, ** p<0.05, *** p<0.01.

Finally, participants who were asked to choose between eight candidates at a time chose attractive candidates and female candidates significantly more often. Black candidates, on the other hand, were chosen significantly less often relative to White candidates. Asians were chosen more often and Latinos less often, though the effect is insignificant (see Table 3.3, column "8"). The rightmost column of the table shows that there is no statistically significant difference between the estimates from two and eight worker treatments, as all the interaction terms are insignificant.

To summarize results of this treatment, we find that hiring chances of females significantly declines in the four-worker treatment, compared to the two-worker treatment. As for race, when the number of available candidates increases hiring bias against Black candidates widens, though we do not have enough statistical power to reject the null hypothesis of equal effects. We interpret this result as only weak evidence in support of our hypothesis stating that increasing the number of available candidates would raise hiring bias[70]. Therefore, discriminatory behavior can be observed when people are overloaded with choices, and it can be driven by implicit preferences.

3.5.5 Additional Analysis

To examine whether the effects of attractiveness differ by the gender of the candidate, we add the interaction of *Female* and *Attractiveness* to our main specification. As the estimation result shows, the coefficient of the interaction term is not statistically significant, suggesting that increasing attractiveness has the same effect for male and female candidates (Table 3.4, column 1).

We also add the predicted number of difficult math questions that candidates got right to our main specification. Column 2 shows that the predicted score is positively and strongly correlated with the hiring decision. Additionally, even after controlling for the predicted score, female and attractive candidates are still hired more often. Although the coefficients of race still have their expected signs, they become statistically insignificant. It appears that, at least for race, differences in hiring can be explained by differences in predicted performance. This is in line with the interpretation of the results above, where we suggest that Asian workers face statistical discrimination.

[70] Appendix C.6 shows the analysis of this section disaggregated by the information provided to employer subjects. Our data suggests no particular treatment manipulation to be the driver of results.

Table 3.4: Estimates of Discrete Choice Model – Additional analyses

Dependent variable: *Choice*	(1)	(2)
Attractiveness	0.08*	0.08**
	(0.05)	(0.04)
Female prop	0.48***	0.54***
	(0.06)	(0.06)
Female * Attractiveness	0.05	
	(0.06)	
Asian prop	0.20**	0.09
	(0.09)	(0.09)
Black prop	-0.17**	-0.08
	(0.07)	(0.08)
Latino prop	-0.12	-0.09
	(0.12)	(0.14)
Prediction		1.07***
		(0.04)
N	9256	9256
Number of clusters	2216	2216
Pseudo R^2	0.025	0.024

Notes: Standard errors are clustered at the subject level. * p<0.1, ** p<0.05, *** p<0.01. Column 1 adds "Female" and "Attractiveness" interaction to the main specification, while column 2 adds "Prediction" as an independent variable.

3.5.6 Are Effects Driven by Other Characteristics?

To find possible underlying mechanisms, as well as to check whether other facial expression characteristics affect hiring decisions, we included those characteristics as explanatory variables. These characteristics were rated by independent coders hired by the Chicago Face Database, and hence were part of our data. Being perceived as angry significantly reduces the workers' chance of being employed (see Table 3.5, column 1), while appearing happy increases that probability, though the effect is statistically insignificant for the latter (see Table 5, column 2). When *Angry* or *Happy* are added as control variables, all explanatory variables that were statistically significant in the original specification remain significant and have the same sign. In column 3, we examine what happens when perceived masculinity is added to the model. Workers that are perceived to be masculine are less likely to be hired, and in this model, the coefficient estimate of gender is no longer statistically significant. Column 4 examines what happens when *Feminine* is added to the model. People that appear feminine are more likely to be hired, and once *Feminine* is added, the effects of attractiveness and gender become much smaller in magnitude and statistically insignificant (see Table 3.5, column 4). This suggests that appearing feminine (or masculine) may be an underlying mechanism for some of our results.[71]

Further examining other facial expressions shows that having a "Dominant" and "Threatening" face significantly reduces hiring probability and a "Trustworthy" face does not have a significant impact on the result (see Table 3.5, columns 5-7). None of these variables affect the significance of the other variables.[72]

Table 3.5: Estimates of Discrete Choice Model – Other characteristics

Dependent variable: *Choice*	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Attractiveness	0.11*** (0.04)	0.13*** (0.04)	0.12*** (0.04)	0.05 (0.05)	0.14*** (0.03)	0.11** (0.04)	0.11*** (0.04)
Female prop	0.50*** (0.06)	0.49*** (0.06)	0.16 (0.16)	-0.00 (0.20)	0.42*** (0.06)	0.48*** (0.06)	0.46*** (0.06)
Asian prop	0.18** (0.07)	0.20** (0.09)	0.17* (0.009)	0.19* (0.09)	0.16* (0.09)	0.17* (0.09)	0.15* (0.09)
Black prop	-0.18** (0.07)	-0.18** (0.07)	-0.13* (0.07)	-0.14* (0.07)	-0.14** (0.07)	-0.19*** (0.07)	-0.17** (0.07)
Latino prop	-0.11 (0.12)	-0.11 (0.12)	-0.11 (0.12)	-0.09 (0.12)	-0.06 (0.13)	-0.10 (0.13)	-0.11 (0.12)
Angry	-0.11** (0.04)						
Happy		-0.04 (0.04)					
Masculine			-0.13** (0.06)				
Feminine				0.18** (0.07)			
Dominant					-0.13*** (0.05)		
Trustworthy						-0.13 (0.09)	
Threatening							-0.14*** (0.05)
N	9256	9256	9256	9256	9256	9256	9256
Number of clusters	2216	2216	2216	2216	2216	2216	2216
Pseudo R^2	0.026	0.025	0.026	0.026	0.026	0.025	0.026

Notes: Standard errors are clustered at the subject level. * p<0.1, ** p<0.05, *** p<0.01. Various characteristics from the Chicago Face Database are added to the main specification in this table.

[71] We have checked results for these "other characteristics" by gender and race of applicants, which are discussed in Appendix G and H.

[72] We also conduct additional robustness checks by removing outliers (e.g. people that responded too quickly or slowly). Results are in the Appendix.

3.5.7 Effect Sizes

We can use odds ratios to compute effect sizes. For example, in the main sample, we find that Blacks are 16 percent less likely to be chosen than Whites ($e^{-0.17}$-1=-0.16). The effect size is smaller than in Bertrand and Mullainathan (2004), who find that African-Americans are 50 percent less likely to receive interview callbacks than Whites. However, the magnitude of discrimination still appears to be sizeable.

The effects of our experimental manipulations are economically meaningful as well. For example, in the two-worker condition, Blacks are chosen around 7 percent less than Whites, but in the eight-worker condition they are chosen around 25 percent less than Whites. Tables with odds ratios can be found in the Appendix C3.

3.6 Potential Interpretations and Implications

Our data do not allow us to pinpoint the underlying mechanism of behavior. However, we give an explanation that is consistent with the unexpected bias in favor of females, and the expected bias we found with regards to attractiveness and race.

This explanation is based on the notion that awareness can reduce biases. Racial bias in professional basketball referees persisted even after a study showed such bias (Price and Wolfers, 2010) but disappeared after extensive media coverage of that study, suggesting that awareness reduced such bias (Pope et al., 2018). Making crowdworkers aware of their own biases reduced their own biases (Hube et al., 2017), and academic promotion committees in scientific fields do not promote more men over women when they believe that gender bias exists (Regner et al., 2019).

It could be that participants in our experiment thought that gender bias was the purpose of this study (being an often-mentioned topic with regards to mathematical performance) and tried to correct for this bias but were overzealous in correcting for it.[73] This happened despite subjects being incentivized to make decisions based on their beliefs. When the prize money was not at stake anymore in the post-experiment survey, the majority of participants stated that math is associated with males. With regards to race and attractiveness, however, subjects were probably unaware of their biases (because

[73] Similarly, Timming et al. (2021) find in their experiment that female opinions were more likely to be acted upon by managers compared to male opinions, the result which authors attribute to social desirability bias – a change in response by subjects in order to be viewed favorably by others.

disparities by race and attractiveness in mathematics are less often mentioned) and probably did not correct for it.

Regarding the effects of our manipulations on information provision, our data suggests that the theory of statistical discrimination can explain positive bias towards attractive candidates and negative bias against Asians. Alternatively, we speculate that showing prior performance at the individual level may have diverted participants' attention towards candidates' past performance. As a result, the effects of all other characteristics disappeared, except for the most salient characteristic (gender). Similarly, displaying information on performance across genders could have made participants more ignorant about other subgroups (e.g., racial minority employees) and associated stereotypes. We emphasize that future research should examine the validity of this explanation.

The gig economy has many stakeholders, and each stakeholder can have multiple objectives (e.g., efficiency, equity). One implication that stands out is that the choice overload can adversely affect certain subgroups. Indeed, there was less evidence of Black candidate discrimination in the two-worker condition than in the four or eight worker conditions[74]. If our results generalize, designers of the online freelance platforms should consider displaying candidates in a way that is less likely to trigger such choice overload. One possible technique that deserves further study is to limit the number of candidates displayed on each page.

A second implication is that provision of additional information to assist in hiring decisions does not have straightforward implications. In our study the relationship between the amount of information provided and discrimination was not monotonic; there was some discrimination when no information on past performance was provided, the most discrimination when information on performance by subgroup (gender) was provided, and the least discrimination when individual-level performance was provided. One potential explanation is that subgroup information could have increased discrimination by reminding people to consider a person's subgroup. Therefore, information should be carefully selected and presented in a way to make sure that subgroups are not unnecessarily adversely affected.

---

[74] Here we should stress that results of the manipulation of number of workers displayed should be taken with caution, as most of the interaction terms are statistically insignificant in the regression.

Lastly, making people aware of their biases can in some cases help disadvantaged groups. If our explanation that the lack of hiring bias against females was due to people being aware of this particular bias is verified by future research, then exploring methods of making people aware of their biases can potentially reduce discrimination in other settings as well.

3.7 Conclusion

We ran an MTurk experiment where we asked participants to make hiring decisions for a mathematically intensive task. Contrary to our expectations, we find that our participants hire females more often than males. On the other hand, racial discrimination occurs largely as expected: Blacks are hired less often than Whites and Asians are hired more often than Whites. Also, attractive candidates are hired more often than less attractive candidates. Moreover, discrimination against Blacks increases as the number of workers a participant can choose from increases, though we cannot reject the null hypothesis of equal treatment across different experimental manipulations. Finally, the relationship between discrimination and information provided to assist hiring decisions is non-monotonic in the amount of information provided.

The immediate takeaway is that since information provision can reduce hiring biases, designers of online freelance platforms can do much to reduce hiring biases. Despite the limitations of our study in pinpointing the exact underlying mechanisms, our findings also serve as a call for further research in this area to determine under what contexts biases in hiring manifest themselves.

Our paper contributes to several streams of literature. Our finding that information provision can affect discrimination is relevant to the human-computer interaction literature as well as the discrimination literature. We also contribute to the choice overload literature by finding weak evidence that choice overload can affect employment decisions, as well as by illustrating how choice overload can affect equity concerns.

There are a few more limitations of our study, on top of those already mentioned. One key limitation is generalizability: our study involved hiring people for mathematically intensive tasks. The kinds of discrimination that appear, as well as the methods of reducing such discrimination, may be different if the nature of the task were changed, particularly if the study was conducted in a field setting. Nonetheless, it is our belief that with persistent study and effort, it is possible to reduce discrimination in many areas, and our paper shows the potential of information provision to decrease discrimination.

Also, our findings may not generalize to settings without photos, such as Amazon MTurk[75]. That said, many online platforms use photos in their worker profiles, such as TaskRabbit, Upwork and Fiverr (to name a few). Even non-gig work marketplaces such as AirBnB, Uber, and Lyft use photos in their worker profiles (and racial discrimination based on photos has been documented in all three of them). Finally, the use of photos in offline resumes is common in European countries such as Germany, as well as China and Japan. Therefore, while it would be useful for future work to examine a setting without photos, we would argue that at the time of writing, an experiment that uses photos is at least as important (if not more important) than an experiment that does not.

[75] It also will not generalize to MTurk because on MTurk, employers (or more precisely, requesters) do not choose workers. Rather, employers set criteria, and anyone who meets them can start the task.

List of References

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

Akerlof, G, Kranton, R. (2000). Economics and Identity. Quarterly Journal of Economics, Vol 115 (3), pp. 715-53.

Altonji, J.G., Zhong, L. (2020). The labor market returns to advanced degrees. NBER working paper 26959.

Antonellis, P.J., Silsbee, R. (2018). Employment Interview Screening: Time to Face the Ink. Journal of Business and Economic Policy, Vol 5(4), pp. 44–53.

Antonellis, P.J., Berry, G., Silsbee, R. (2017). Employment Interview Screening: Is The Ink Worth It? Global Journal of Human Resource Management, Vol 5 (2), pp. 38-53.

Arceo-Gomez, E.O., Campos-Vazquez, R.M. (2014). Race and Marriage in the Labor Market: A Discrimination Correspondence Study in a Developing Country. American Economic Review, Vol 104(5), pp. 376–380.

Aslam, A., Owen, C. (2013). Fashions change but tattoos are forever: time to regret. British Journal of Dermatology, Vol 169 (6), pp. 1364-1366.

Baert, S. (2018). Hiring Discrimination: An Overview of (Almost) All Correspondence Experiments Since 2005 In: Gaddis S. (eds) Audit Studies: Behind the Scenes with Theory, Method, and Nuance. Vol 14, pp. 63–77. Springer, Cham.

Baert, S., Balcaen, P. (2013). The Impact of Military Work Experience on Later Hiring Chances in the Civilian Labour Market. Evidence from a Field Experiment. Economics: The OpenAccess, Open-Assessment E-Journal, 7 (2013-37): 1–17.

Baert, S., De Pauw, A.S., Deschacht, N. (2016a). Do Employer Preferences Contribute to Sticky Floors? ILR Review Vol 69, pp. 714–736.

Bartoš, V., M. Bauer, J. Chytilová, Matejka, F. (2016). Attention Discrimination: Theory and Field Experiments with Monitoring Information Acquisition. American Economic Review, Vol 106(6), pp. 1437–1475.

Baumann, C., Timming, A. R., Gollan, P. J. (2016). Taboo tattoos? A study of the gendered effects of body art on consumers' attitudes toward visibly tattooed front line staff. Journal of Retailing and Consumer Services, Vol 29, pp. 31–39.

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

Becker, G. (1971). The Economics of Discrimination. Second edition, University of Chicago Press.

Becker, S. O., Fernandes, A., Weichselbaumer, D. (2019). Discrimination in Hiring Based on Potential and Realized Fertility: Evidence from a Large-Scale Field Experiment. Labour Economics Vol 59, pp. 139–152.

Ben-Akiva, M., Bierlaire, M. (1999). Handbook of Transportation Science. Vol. 23. Springer, Boston, MA, Chapter Discrete Choice Methods and their Applications to Short Term Travel Decisions, 00. 5–33.

Ben-Akiva, M., Boccara, B. (1995). Discrete choice models with latent choice sets. International Journal of Research in Marketing Vol 12 (1), pp. 9–24.

Bertrand, M., Duflo, E. (2017). Field Experiments on Discrimination. Handbook of Field Experiments, Vol 1.

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

Bohnet, I., van Geen, A., Bazerman, M. (2016). When Performance Trumps Gender Bias: Joint vs. Separate Evaluation. Management Science Vol 62 (5), pp. 1225–1531.

Bohren, I., Imas, I., Rosenberg, M. (2019). The Dynamics of Discrimination: Theory and Evidence. American Economic Review Vol 109 (10), pp. 3395-3436.

Bordalo, P., Coffman, K., Gennaioli, N., Shleifer, A. (2016). Stereotypes. The Quarterly Journal of Economics, Vol 131 (4), pp. 1753-1794.

Brallier, S., Maguire, K., Smith, D., Palm, L. (2011). Visible Tattoos and Employment in the Restaurant Service Industry. International Journal of Business and Social Science, Vol 2 (6), pp. 72-76.

Cameron, C., Miller, D. (2015). A Practitioner's Guide to Cluster-Robust Inference. Journal of Human Resources Vol 50 (2), pp. 317–372.

Carlsson, M., Rooth, D. (2012). Revealing taste-based discrimination in hiring: A correspondence testing experiment with geographic variation. Applied Economics Letters, Vol 19(18), pp. 1861–1864.

Cherneva, A., Böckenholt, U., Goodman, J. (2015). Choice overload: A conceptual review and meta-analysis. Journal of Consumer Psychology Vol 25 (2), pp. 333–358.

Coffman, K., Exley, C., Niederle, M. (2019). When gender discrimination is not about gender. Boston: Harvard Business School.

Cohn, A., Marechal, M.A., Schneider, F., Weber, R.A. (2017). Job History, Non-Cognitive Skills,

Dillingh, R., Kooreman, P., Potters, J. (2019). Tattoos, Lifestyle, and the Labor Market. Labour, Vol 34(2), pp. 191–214.

Dolaec, J.L., Stein, L.C..D (2013). The Visible Hand: Race and Online Market Outcomes. The Economic Journal, Vol 123(572), pp. F469–F492.

Edelman, B., Luca, M., Svirsky, D. (2017). Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment. American Economic Journal: Applied Economics Vol 9 (2), pp. 1–22.

Ellemers, N. (2018). Gender stereotypes. Annual review of psychology Vol 69, pp. 275–298.

Equality Challenge Unit. (2009). The Experience of Lesbian, Gay, Bisexual and Trans Staff and Students in Higher Education. ECU: London.

Federal Government Commissioner for the New Federal States. (2019). Annual Report of the Federal Government on the Status of German Unity. Federal Ministry for Economic Affairs and Energy, Berlin.

French, M., Maclean, J.C., Robins, P., Sayed, B., Shiferaw, L. (2016). Tattoos, Employment, and Labor Market Earnings: Is There a Link in the Ink? Southern Economic Journal, 82 (4), pp. 1212-1246.

French, M., Mortensen, K., Timming, A. (2019). Are tattoos associated with employment and wage discrimination? Analyzing the relationships between body art and labor market outcomes. Human Relations, Vol 72(5), pp. 962–987.

Fryer, R., Jackson, M. (2008). A Categorical Model of Cognition and Biased Decision Making. The B.E. Journal of Theoretical Economics, Vol. 8 Iss 1, Article 6.

Gneezy, U., List, J., Price, M. (2012). Toward an understanding of why people discriminate: Evidence from a series of natural field experiments. NBER Working paper 17855

Goldberg, A., Mourinho, D., Kulke, U. (1996). Labour Market Discrimination against Foreign Workers in Germany. International Migration Papers.

Greenwald, A.G., McGhee, D.E., Schwartz, J.L.K. (1998). Measuring individual differences in implicit cognition: the implicit association test. Journal of Personality and Social Psychology, Vol 74 (6), pp. 1464-1080.

Guéguen N. (2013). Effects of a tattoo on men's behavior and attitudes towards women: An experimental field study. Arch Sex Behav. Vol 42(8), pp. 1517-1524.

Guryan, J., Kofi Charles, K. (2013). Taste-based or statistical discrimination: the economics of discrimination returns to its roots. The Economic Journal 123 (572), F417–F432.

Hamermesh, D., Biddle, J. (1994). Beauty and the Labor Market. American Economic Review, Vol. 84, No. 5, pp. 1174-1194

Hannak, A., Wagner, C., Garcia, D., Mislove, A., Strohmaier, M., Wilson, C. (2017). Bias in online freelance marketplaces: Evidence from Taskrabbit and Fiverr.

Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (2017), pp. 1914–1933.

Harper, B. (2000). Beauty, Stature and the labor market: A British cohort study. Oxford Bulletin of Economics and Statistics, 62, pp. 771-800.

Harrison, G., List, J. (2004). Field Experiments. Journal of Economic Literature, Vol 42 (4), pp. 1009-1055.

Hawkes, D., Senn, C. Y., Thorn, C. (2004). Factors that influence attitudes toward women with tattoos. Sex Roles, Vol 50, pp. 593–604.

Heckman, J., Siegelman, P. (1993). The Urban Institute Audit Studies: Their Methods and Findings. In Clear and Convincing Evidence: Measurement of Discrimination in America, ed. Fix and Struyk, pp. 187–258. Washington, D.C.: The Urban Institute Press.

Hube, C., Fetahu, B., Gadiraju, U. (2019). Understanding and Mitigating Worker Biases in the Crowdsourced Collection of Subjective Judgments. CHI 1, 1 (2019), 407.

Iyengar, S., Kamenica, E. (2010). Choice proliferation, simplicity seeking, and asset allocation. Journal of Public Economics Vol 94 (7), pp. 530–539.

Iyengar, S., Lepper, M. (2000). When Choice is Demotivating: Can One Desire Too Much of a Good Thing? Journal of Personality and Social Psychology Vol 79 (6), pp. 995–1006.

Jackman, P., Hawkins, R., Burke, S., Swann, C., Crust, L. (2020). The psychology of mountaineering: a systematic review. International Review of Sport and Exercise Psychology. Advanced online publication. https://www.tandfonline.com/doi/full/10.1080/1750984X.2020.1824242

Jenq, C., Pan, J., Theseira, W. (2015). Beauty, weight, and skin color in charitable giving. Journal of Economic Behavior and Organization Vol 119, pp. 234–253.

Jibuti, D. (2018). Discrimination against Workers with Visible Tattoos: Experimental Evidence from Germany. CERGE-EI Working Paper Series No. 628

Jowell, R., Prescott-Clarke, P. (1970). Racial Discrimination and White-Collar Workers in Britain. Race & Class, Vol 11 (4), pp. 397-417.

Kaas, L., Manger, C. (2012). Ethnic Discrimination in Germany's Labour Market: A Field Experiment. German Economic Review, Vol 13, pp. 1-20.

Kahneman, D. (2011). Thinking fast and slow. Farrar, Straus and Giroux.

Kamenica, E. (2008). Contextual Inference in Markets: On the Informational Content of Product Lines. American Economic Review Vol 98 (5), pp. 2127-2149.

Katuscak, P., Kraft, P. (2013). (Not) Just another pretty face: Experimental evidence for discrimination in recruitment. CERGE-EI working paper.

Lancsar, E., Louviere, J. (2008). Conducting Discrete Choice Experiments to Inform Healthcare Decision Making. PharmacoEconomics Vol 26 (8), pp. 661–677.

Lee, M.K., Kiesler, S., Forlizzi, J. (2011). Mining behavioral economics to design persuasive technology for healthy choices. CHI 1, 1 (2011), 325–334.

Levitt, S., List, J. (2007). What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World? Journal of Economic Perspectives, Vol 21, pp. 153–174.

List, J., Shaikh, A., Xu, X. (2016). Multiple hypothesis testing in experimental economics. NBER Working paper 21875.

List, J. (2008). Introduction to field experiments in economics with applications to the economics of charity. Experimental Economics, Vol 11 (3), pp. 203–212.

List, J. (2006). Field Experiments: A Bridge between Lab and Naturally Occurring Data. Advances in Economic Analysis & Policy, Vol 6 (2), Article 8.

Loewenstein, G. (1999). Because it is there: the challenge of mountaineering for utility theory? Kyklos, Vol 52 (3), pp. 315-344.

Mobius, M., Rosenblat, T. (2006). Why Beauty Matters. American Economic Review Vol 96 (1), pp. 222–235.

Morton, F.S., Zettelmeyer, F., Silva-Risso, J. (2003). Consumer Information and Discrimination: Does the Internet Affect the Pricing of New Cars to Women and Minorities? Quantitative Marketing and Economics Vol 1 (1), pp. 65–92.

Mortensen, K., French, M. T., Timming, A. (2019). Are tattoos associated with negative health-related outcomes and risky behaviors? International Journal of Dermatology, Vol 58 (7), pp. 816–824.

Neumark, D. (2012). Detecting Discrimination in Audit and Correspondence Studies. Journal of Human Resources, Vol 47(4), pp. 1128–1157.

Neumark, D. (2018). Experimental Research on Labor Market Discrimination. Journal of Economic Literature, Vol 56(3), pp. 799–866.

Neumark, D. (2020). Age Discrimination in Hiring: Evidence from Age-blind vs. Non-age-blind Hiring Procedures. NBER working paper 26623.

Neumark, D., Burn, I., Button, P. (2019). Is It Harder for Older Workers to Find Jobs? New and Improved Evidence from a Field Experiment. Journal of Political Economy Vol 127(2), pp. 922–970.

Nevo, A. (2000). A Practitioner's Guide to Estimation of Random-Coefficients Logit Models of Demand. Journal of Economics and Management Strategy Vol 9 (4), pp. 513–548.

Nüß, P., Penny, M. (2019). Now You See Me! Ethnic Discrimination in the Market for Apprenticeships. Unpublished manuscript.

Pew Research Center (2010). Millennials: A Portrait of Generation Next. Report.

Phelps, S. (1972). The Statistical Theory of Racism and Sexism. American Economic Review, Vol 62 (4), pp.659-661.

Pope, D., Price, J., Wolfers, J. (2018). Awareness Reduces Racial Bias. Management Science Vol 64 (11), pp. 4967–5460.

Pope, D., Sydnor, J. (2011). What's in a Picture? Evidence of Discrimination from Prosper.com. Journal of Human Resources Vol 46 (1), pp. 53–92.

Price, J., Wolfers, J. (2010). Racial Discrimination Among NBA Referees. Quarterly Journal of Economics Vol 125 (4), pp. 859–1887.

Quillian, L., Pager, D., Hexel, O., Midtbøen, A.H. (2017). Meta-analysis of field experiments shows no change in racial discrimination in hiring over time. PNAS 1, 1, 1–6.

Regner, I., Thinus-Blanc, C., Netter, A., Schmader, T., Huguet, P. (2019). Committees with implicit biases promote fewer women when they do not believe gender bias exists. Nature Human Behaviour Vol 3, pp. 1171–1179.

Riach, P., Rich, J. (1987). Testing for Sexual Discrimination in the Labour Market. Australian Economic Papers, Vol 26, pp. 165-178.

Riach, P., Rich, J. (2002). Field Experiments of Discrimination in the Market Place. The Economic Journal, Vol 112, pp. 480-518.

Riach, P., Rich, J. (2006). An Experimental Investigation of Sexual Discrimination in Hiring in the English Labor Market. Advances in Economic Analysis and Policy, Vol 6 (2), pp. 1-20.

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

Ruffle, B., Shtudiner, Z. (2015). Are Good-Looking People More Employable? Management Science, Vol 61(8), pp. 1760-1776.

Ruffle, B.J., Wilson, A. (2018). The truth about tattoos. Economics Letters, Vol 172, pp. 143–147.

Ruffle, B.J., Wilson, A. (2019). Tat will tell: Tattoos and time preferences. Journal of Economic Behavior & Organization, Vol 166, pp. 566–585.

Smith, A. (2016). Gig Work, Online Selling and Home Sharing.

Stirn, A., Hinz, A., Brahler, E. (2006). Prevalence of tattooing and body piercing in Germany and perception of health, mental disorders, and sensation seeking among tattooed and bodypierced individuals. Journal of Psychosomatic Research, Vol 60, pp. 531-534.

Swami, V., Gaughan, H., Tran, U. S., Kuhlmann, T., Stieger, S., Voracek, M. (2015). Are tattooed adults really more aggressive and rebellious than those without tattoos? Body Image, Vol 15, pp. 149–152.

Swami, V., Tran, U. S., Kuhlmann, T., Stieger, S., Gaughan, H., Voracek, M. (2016). More similar than different: Tattooed adults are only slightly more impulsive and willing to take risks than Non-tattooed adults. Personality and Individual Differences, Vol 88, pp. 40–44.

Swanger, N. (2006). Visible body modification (VBM): Evidence from human resource managers and recruiters and the effects on employment. International Journal of Hospitality Management, Vol 25, pp. 154-158.

Tajfel, H., Turner, J. (1979). An Integrative Theory of Intergroup Conflict. The Social Psychology of Intergroup Relations, pp 33-47.

Tang, S., Zhang, X., Cryan, J., Metzger, M., Zheng, H., Zhao, B.Y. (2017). Gender bias in the job market: A longitudinal analysis. Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing 1, 2.

Thaler, R., Sunstein, C. (2009). Nudge: Improving Decisions About Health, Wealth, and Happiness. Penguin.

Timming, A. (2017). Body art as branded labour: At the intersection of employee selection and relationship marketing. Human Relations, Vol 70 (9), pp. 1041-1063.

Timming, A. (2015). Visible tattoos in the service sector: a new challenge to recruitment and selection. Work, employment and society, Vol 29 (1), pp. 60-78.

Timming, A., Baumann, C., Gollan P. (2021). Employee voice and perceived attractiveness: are less attractive employees ignored in the workplace? Journal of Participation and Employee Ownership, Vol 4 (2), pp. 26-41.

Timming, A., Perrett, D. (2017). An experimental study of the effects of tattoo genre on perceived trustworthiness: Not all tattoos are created equal. Journal of Trust Research, Vol 7 (2), pp. 115-128.

Timming, A., Nickson, D., Re, D., Perrett, D. (2017). What Do You Think of My Ink? Assessing the effects of body art on employment chances. Human Resource Management, Vol 56 (1), pp. 133-149.

U.S. Bureau of Labor Statistics, (2020). Table 11: Employed Persons by Detailed Occupation, Sex, Race, and Hispanic or Latino Ethnicity, 2019. Current

Population Survey, Household Data Annual Averages 2019. Available from https://www.bls.gov/cps/cpsaat11.htm

Walzer, A., Sanjurjo, P. (2016). Media and contemporary tattoo. Communication & Society, Vol 29(1), pp. 69–81.

Weichselbaumer, D. (2019). Multiple Discrimination against Female Immigrants Wearing Headscarves. Industrial and Labor Relations Review, Vol 73(3), pp. 600–627.

Weichselbaumer, D. (2017). Discrimination against migrant job applications in Austria: An experimental Study. German Economic Review, Vol. 18 (2), pp. 237-265.

Wooldridge, J. (2019). Introductory Econometrics: A Modern Approach. South-Western College Pub.

Zestcott, C.A., Tompkins, T.L., Kozak Williams, M., Livesay, K., Chan, K.L. (2017). What do you think about ink? An examination of implicit and explicit attitudes toward tattooed individuals. The Journal of Social Psychology, Vol 158(1), pp. 7–22.

List of Appendices

A.1 Results of Applicant Attractiveness Survey

To rank the experiment participants according to their attractiveness, I conducted a survey. The survey included pictures of the participants and asked subjects to assign them to a category - above average looking (compared to the same age/gender group), average looking and below average looking. The survey was sent to volunteer students of different nationalities studying in various universities in Prague, the Czech Republic. In total 35 subjects responded to the survey. There was a total of 10 applicant pictures in the survey, 4 female and 6 males. After the survey, I chose a picture of one female and male participant from each category. As Table A.7 shows, there is a consensus about participants' attractiveness among the students surveyed.

Table A.7: Attractiveness evaluation of experiment participants

Participant #	**Attractiveness category**		
	Above average looking	*Average looking*	*Below average looking*
1	0	40	60
2	85.8	8.6	5.6
3	31.4	48.6	20
4	20	62.9	17.1
5	20	31.4	48.6
6	80	14.1	5.9

Notes: Numbers are percentages of votes in the survey

I wanted to make sure that applicants are not rejected because of characteristics other than a visible tattoo, for ethnicity, for example, so I wanted candidates to have a "German look". Ideally one could use pictures of German people, although in my case it was not feasible, as only one participant is from Germany. For this reason, I needed to make sure that participants' perceived nationality was similar in the treatment and control group. In the first survey described above, in addition to perceived attractiveness, I asked respondents to state (their perceived) nationality of the person depicted on the pictures.[76] Alongside this survey I created another survey, this time using photos with tattoos, and asked another set of participants to state the perceived nationality of the person in the picture. 26 volunteers completed the survey. Table A.8 shows the top three nationalities indicated by volunteers (with respective percentages). As the table shows, there is no difference in the perceived nationality of applicants with and without tattoos. This ensures

[76] In that survey participants did not have tattoo.

that applications in the treatment and control group will not be treated as different nationals, which may complicate the results.

Table A.8: Perceived nationality of experiment participants

Without Tattoo	Participant #	With Tattoo
Top 3 nationalities		Top 3 nationalities
German - 31%; Czech - 26%; British - 22%	1	Czech - 42%; German - 27%; British - 15%
German - 29%; American - 29%; British - 17%	2	German - 35%; Czech - 23%; British - 23%
British - 67%; American - 17%; German - 6%	3	British - 46%; German - 27%; American - 23%
Czech - 26%; American - 17%; German - 9%	4	American - 31%; Czech - 23%; German - 8%
Czech - 31%; American - 29%; German - 20%	5	American - 31%; British - 31%; Czech - 15%
American - 29%; Czech - 26%; German - 23%	6	American - 35%; British - 31%; Czech - 23%

A.2 Randomization Check

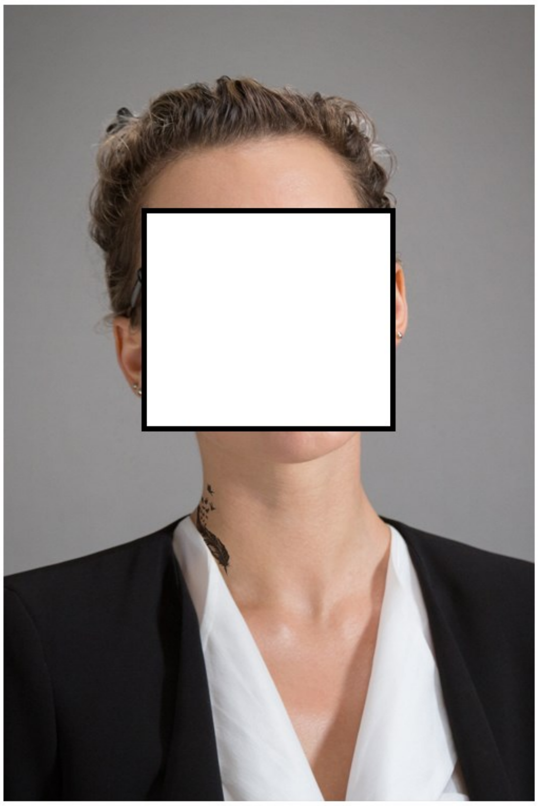
As I sent only one application to one employer, I needed to ensure that firms and jobs were similar in the treatment and control group in terms of all controllable characteristics. In the paper I presented evidence that in terms of a firm's characteristics the sample is balanced. Here I do the same exercise for regions. I test whether regions of the country are similarly represented in the treatment and control group. Table A.9 shows balanced check results for regions. None of the differences are statistically significant, meaning that the randomization ensures the treatment and the control groups are similar in terms of controllable characteristics. Thus, I can rule out that any differential treatment of tattooed applicants is related to firm characteristics and/or to region- specific factors. Therefore, I argue that any difference in callback rates between the treatment and the control group should be due to the treatment itself.

Table A.9: Randomization check – regions

Region	*Non-Tattooed*	*Tattooed*	*P-value*
Baden-Württemberg	0.09 (0.29)	0.09 (0.29)	0.89
Bavaria	0.19 (0.39)	0.19 (0.39)	0.84
Berlin	0.10 (0.31)	0.10 (0.30)	0.98
Brandenburg	0.00 (0.05)	0.01 (0.07)	0.58
Bremen	0.02 (0.14)	0.03 (0.16)	0.68
Hamburg	0.08 (0.27)	0.08 (0.28)	0.89
Hesse	0.14 (0.35)	0.14 (0.35)	0.94
Lower Saxony	0.03 (0.18)	0.04 (0.19)	0.63
Mecklenburg-Vorpommern	0.00 (0.05)	0.01 (0.07)	0.58
North Rhine-Westphalia	0.22 (0.41)	0.22 (0.41)	0.90
Rhineland-Palatinate	0.01 (0.10)	0.01 (0.09)	0.67
Saarland	0.00 (0.05)	0.00 (0.05)	0.98
Saxony	0.05 (0.01)	0.04 (0.01)	0.34
Saxony-Anhalt	0.01 (0.07)	0.01 (0.09)	0.43
Schleswig-Holstein	0.03 (0.17)	0.02 (0.15)	0.60
Thuringia	0.01 (0.09)	0.01 (0.09)	0.74
N	385	397	

Notes: The table shows mean comparison of regions across treatment (tattooed) and the control (non-tattooed) groups. Standard deviations are in parenthesis. Column 3 shows p-values of the hypothesis of equal means.

A.3 Pictures Used in the Experiment



A.4 Robustness Check – Probit Model Estimates

To perform the robustness of the Linear Probability Model (LPM) used in the main text I performed the same analysis using Probit model. Tables below confirm that the Probit model produces results that are qualitatively same the LPM model results.

Table A.10: Estimates of the Probit Model – Firm Characteristics

Dependent variable: Callback	(1)	(2)	(3)	(4)	(5)	(6)
Visible Tattoo	-0.13***††† (0.03)	-0.13***††† (0.03)	-0.12** (0.05)	-0.10***† (0.04)	-0.18** (0.07)	-0.15* (0.08)
Male			-0.12** (0.05)			
Visible Tattoo * Male			-0.01 (0.06)			
International firm				-0.04 (0.05)		
Visible Tattoo * International firm				-0.08 (0.07)		
Visible Tattoo * West Germany					0.07 (0.08)	
Urban area						-0.06 (0.07)
Visible Tattoo * Urban area						0.03 (0.09)
Constant	0.30***††† (0.02)	0.30***††† (0.02)	0.30***††† (0.02)	0.30***††† (0.02)	0.30***††† (0.02)	0.30***††† (0.02)
Monthly and regional dummies	*N*	*Y*	*Y*	*Y*	*Y*	*Y*
Control variables	*N*	*N*	*Y*	*Y*	*Y*	*Y*
Pseudo R^2	0.02	0.07	0.11	0.10	0.10	0.10
N	782	782	782	782	782	782

Notes: The table shows average marginal effects of the Probit model. Robust standard errors in parentheses. Columns 2-6 include monthly and regional dummies. In columns 3-6, I control for firm characteristics including age, size, number of job advertisements and whether the location of the job is in an urban area. * p<0.1, ** p<0.05, *** p<0.01. † - significance level with multiple hypothesis testing adjusted.

Table A.11: Estimates of the Probit Model – Job requirements

Dependent variable: Callback	(1)	(2)	(3)	(4)
Visible Tattoo	-0.13***††† (0.03)	-0.13***†† (0.04)	-0.15***††† (0.03)	-0.13*** (0.04)
Front office		0.04 (0.04)		
Visible Tattoo * Front office		0.02 (0.06)		
Appearance requirement			-0.04 (0.05)	
Visible Tattoo * Appearance requirement			0.08 (0.07)	
Teamwork requirement				-0.08 (0.05)
Visible Tattoo * Teamwork requirement				0.05 (0.07)
Constant	0.29***††† (0.02)	0.30***††† (0.02)	0.28***††† (0.02)	0.28***††† (0.02)
Monthly and regional dummies	*Y*	*Y*	*Y*	*Y*
Control variables	*N*	*Y*	*Y*	*Y*
R^2	0.06	0.09	0. 09	0.11
N	782	782	782	782

Notes: The table shows average marginal effects of the Probit model. Robust standard errors in parentheses. Columns 2-4 include monthly and regional dummies. In columns 2-4, I control for firm characteristics including age, size, number of job advertisements and whether the location of the job is in an urban area. * p<0.1, ** p<0.05, *** p<0.01. † - significance level with multiple hypothesis testing adjusted.

Table A.12: Estimates of the Probit Model – Channels of statistical discrimination

Dependent variable: Callback	(1)	(2)	(3)
Visible Tattoo	-0.13***†††	-0.12***††	-0.15***†††
	(0.03)	(0.04)	(0.04)
Reference signal		-0.03	
		(0.05)	
Visible Tattoo * Reference signal		-0.02	
		(0.07)	
Group membership			-0.00
			(0.05)
Visible Tattoo * Group membership			0.09
			(0.06)
Constant	0.29***†††	0.30***†††	0.30***†††
	(0.02)	(0.02)	(0.02)
Monthly and regional dummies	*Y*	*Y*	*Y*
Control variables	*N*	*Y*	*Y*
R^2	0. 06	0.08	0.11
N	782	782	782

Notes: The table shows average marginal effects of the Probit model. Robust standard errors in parentheses. All specifications control for monthly and regional dummies and firm characteristics including age, size, number of job advertisements and whether the location of the job is in an urban area. * p<0.1, ** p<0.05, *** p<0.01. † - significance level with multiple hypothesis testing adjusted.

B.1 Robustness Check – Probit Model Estimates

Table B.4: Estimates of the Probit model

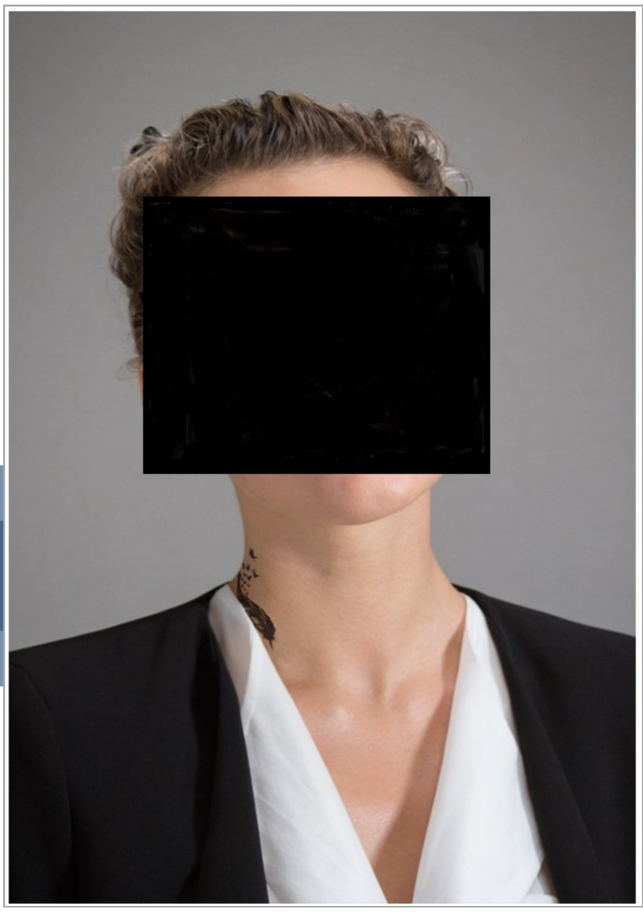
Dependent variable: Callback	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Visible Tattoo	-0.09*** (0.03)	-0.07** (0.03)	-0.09** (0.05)	-0.06 (0.04)	-0.01 (0.09)	-0.10 (0.07)	-0.03 (0.06)
Male			-0.07* (0.04)				
Visible Tattoo * Male			0.03 (0.06)				
High skill				0.05 (0.04)			
Visible Tattoo * High skill				-0.03 (0.06)			
Visible Tattoo * West Germany					-0.07 (0.09)		
Urban area						-0.02 (0.06)	
Visible Tattoo * Urban area						0.04 (0.08)	
Small firm							0.09 (0.08)
Visible Tattoo * Small firm							-0.13 (0.09)
Medium firm							-0.02 (0.07)
Visible Tattoo * Medium firm							-0.01 (0.08)
Constant	0.31*** (0.02)	0.31*** (0.02)	0.31*** (0.02)	0.31*** (0.02)	0.31*** (0.02)	0.31*** (0.02)	0.31*** (0.02)
Monthly and regional dummies	*N*	*Y*	*Y*	*Y*	*Y*	*Y*	*Y*
Control variables	*N*	*Y*	*Y*	*Y*	*Y*	*Y*	*Y*
Pseudo R^2	0.007	0.055	0.059	0.055	0.056	0.056	0.060
N	800	799	799	799	799	799	799

Notes: The table shows average marginal effects of the Probit model. Robust standard errors in parentheses. Columns 2-7 include monthly and regional dummies. In columns 2-7, I control for firm characteristics including age, size, \# of job advertisements, \# of required programs, gender of HR contact, whether the position includes "senior" in the title or requires teamwork, whether the location of the job is in an urban area and whether the firm is international. * p<0.1, ** p<0.05, *** p<0.01.

B.2 Sample Application of High Skilled Female Applicant

Name Surname

Software Engineer



Attachments:
Cover Letter
Resume

Street name ##| Postcode City
xxxxxx@xxxmail.xxxx | +49 XX XXXXXXX

Company name
Address
XXXXX City

Name Surname
Street name XX
Postcode Coty

Date of applying

Application for the position of *POSITION NAME*

Dear *Mrs/Mr Surname*,

The creation of dimensioning and design software for XXXXXX, XXXXX XXXXX XXXXX are very promising areas of responsibility for me. On the position I will benefit from the experience I was able to gain in the context of my current work as a software developer in the area XXXX XXXXX for XXXX in *CITY*. In the course of this activity I have acquired profound know-how in software development, especially in XXXXX XXXXX XXXXX & XXXXX XXXXX XXXXX and the conception of database solutions. In doing so, I demonstrated my profound knowledge as XXXXX XXXXX XXXXX at *UNIVERSITY NAME*, *CITY* and my pronounced analytical and conceptual skills. In XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX and XXXXX XXXXX XXXXX I also use my independent and goal-oriented way of working optimally.

During this activity I built on my knowledge of XXXXX XXXXX XXXXX, which I worked XXXXX XXXXX XXXXX for the *COMPANY NAME*. On this position, I brought in my experienced IT knowledge in the field of software development and participated actively and competently in a variety of IT projects. The focus was on the implementation and optimization of sophisticated IT applications for banks and financial service providers based on Java and C #.

You can expect from me an extensive knowledge of the programming language such as C # / VB.NET and XXXXX XXXXX XXXXX, which is particularly relevant for the position I am applying.

Since your position offers me a very diverse XXXXX XXXXX XXXXX, I would like to take the opportunity with you and convince you as soon as possible with my high IT expertise, taking into account my notice period of 2 weeks. My annual salary expectation for this position is *AMOUNT* Euro p.a.

I would like to start working with you as soon as possible and, of course, I am also willing to change my place of residence for this exciting job.

I look forward to supporting your team in *CITY* as soon as possible with my high level of commitment and I am looking forward to your feedback.

Kind regards

RESUME

NAME SURNAME

Street name, ## · Postcode City · xxxx@xxxmail.xxxx | +49 XXXX XXXXXXX

PERSONAL DATA

Nationality: German
Date of birth: DD MMMMM YYYY

PROFESSIONAL EXPERIENCE

MM/YYYY – today
COMPANY NAME – CITY
POSITION TITLE

- Transformation of existin XXXXX XXXXX XXXXX
- XXXXX XXXXX XXXXX MS SQL Server & Web-applications
- Development and XXXXX XXXXX XXXXX
- XXXXX XXXXX XXXXX of proposed solutions and responsibility for timely implementation

MM/YYYY – MM/YYYY
COMPANY NAME – CITY
POSITION TITLE

- Participation in XXXXX XXXXX XXXXX
- Realization and optimization of demanding IT applications XXXXX XXXXX XXXXX Java or C#
- XXXXX XXXXX XXXXX, desing and implementation before testing
- XXXXX XXXXX XXXXX of database solutions

APPRENTICESHIP

MM/YYYY – MM/YYYY
COMPANY NAME – CITY
IT internship POSITION TITLE

- XXXXX XXXXX XXXXX applications with .NET and C#
- XXXXX XXXXX XXXXX with WPF and ASP.NET MVC
- XXXXX XXXXX XXXXX of business logic with C#, Webservices and MS-SQL

MM/YYYY – MM/YYYY
COMPANY NAME – CITY
Voluntary internship during the semester break, IT

- Active participation in IT projects and XXXXX XXXXX XXXXX
- Creation, XXXXX XXXXX XXXXX in Java and JavaScript
- XXXXX XXXXX XXXXX in business processes and services as well as the implementation of new and modification of existing applications

EDUCATION

MM/YYYY – MM/YYYY
UNIVERSITY NAME, CITY

Diploma: Master of Science in IT
GPA: 1.7

MM/YYYY – MM/YYYY
UNIVERSITY NAME, CITY
Diploma: Bachelor of Science in IT
GPA: 1.3

HIGH SCHOOL EDUCATION

MM/YYYY – MM/YYYY
HIGH SCHOOL NAME, CITY
Diploma: General University Entrance Qualification
GPA: 1.3

IT- KNOWLEDGE

Programming languages: Java, J2EE, C++, XXXXX, Delphi, PHP, XXXXX, MS SQL XML, HTML, VB.NET

Experiences in SPSS, Matlab with XXXXX, WinCC, Step7/Simatic,

Very good knowledge of MSOffice, XXXXX, Unix/Linux

LANGUAGES

German (native)
English (advanced)

SIGNATURE

CITY, DD.MM.YYYY

C.1 Discrete Choice Model

The mathematical equations for the discrete choice model are as follows.[77] Decision makers observe utility of option *i* as $Utility\_i = Value\_i + \varepsilon_i$, where ε_i is the error term. They then maximize their own utility by choosing the option with the highest utility (or if there is a tie, randomly choosing between items with the highest utility). If errors are Type I Extreme Value distributed[78], then the probability that an option will be chosen is calculated as follows:

$$P(\text{option i chosen}) = \frac{e^{Value_i}}{\sum_{j=1}^{N} e^{Value_j}}$$

In other words, the probability that an option is chosen is its exponentiated value, divided by the sum of exponentiated values of all options. This particular form of a discrete choice model is known as a conditional logit model.

C.2 Randomization Check

The two tables below replicate Tables 3.2 and 3.3 respectively, but with the outliers dropped. A participant is defined as an outlier if they answered too slowly/fast given the number of workers they had to choose from (more than 2.5 standard deviations below or above the mean for the number of workers they had to choose from), and/or if they failed attention checks. Specifically, 4 subjects were dropped in the 2-worker condition since they spent more than 144 seconds (more than 2.5 standard deviations above the mean). For the same reason, 3 subjects were dropped in the 8-worker condition as it took them more than 127 seconds to complete the survey. No subject was dropped in 4 worker conditions or because they completed the survey too fast. As an attention check, we asked participants what type of questions potential employees answered. 22 subjects answered either "Liberal Arts" or "Other", while the correct answer was "Math"/"Science". Therefore, we dropped those 22 subjects who failed the attention check question. After dropping the outliers, 157 subjects remain and are used as the subject pool in the robustness checks.

[77] Value is computed as described in the main text.
[78] This is the standard assumption made by the literatures in various disciplines that use discrete choice modelling, arbitrary as it may be. See the references we cited in the main text.

Table C.6 confirms that the qualitative results remain unchanged when we drop outlier observations: All explanatory variables that were statistically significant in Table 2 remain statistically significant and have the same sign. The estimates are also largely similar in magnitude, with only a few exceptions. For example, the estimate of *Female* is reduced to 0.20 after dropping outliers (it was 0.36 in Table 3.2), and the estimate of *Black* and subgroup information treatment interaction is also lower compared to Table 3.2 values (-0.58 vs -0.41) (leftmost column of Table C.6). Similarly, the qualitative results remain the same as in Table 3.3 when outliers are dropped from analysis (Table C.7).

Table C.6: Replication of Table 3.2 models with outliers dropped

Dependent variable: *Choice*	Full sample	Prior performance shown			
		none	*individual*	*subgroup*	*All*
Attractiveness	0.14***	0.25***	0.02	0.17***	0.25***
	(0.04)	(0.06)	(0.06)	(0.06)	(0.06)
Attractiveness * Individual info treatment					-0.23***
					(0.09)
Attractiveness * Subgroup info treatment					-0.09
					(0.09)
Female prop	0.38***	0.20**	0.55***	0.40***	0.20**
	(0.06)	(0.10)	(0.10)	(0.11)	(0.10)
Female prop * Individual info treatment					0.35**
					(0.15)
Female prop * Subgroup info treatment					0.20
					(0.15)
Asian prop	0.28**	0.34**	-0.01	0.44***	0.34**
	(0.09)	(0.16)	(0.17)	(0.15)	(0.16)
Asian prop * Individual info treatment					-0.35
					(0.24)
Asian prop * Subgroup info treatment					0.10
					(0.22)
Black prop	-0.14*	0.12	-0.10	-0.46***	0.12
	(0.08)	(0.13)	(0.14)	(0.14)	(0.13)
Black prop * Individual info treatment					-0.22
					(0.19)
Black prop * Subgroup info treatment					-0.58***
					(0.19)
Latino prop	-0.07	0.18	0.01	-0.44*	0.18
	(0.14)	(0.22)	(0.25)	(0.25)	(0.22)
Latino prop * Individual info treatment					-0.17
					(0.33)
Latino prop * Subgroup info treatment					-0.62*
					(0.33)
N	7726	2770	2296	2660	7726
Number of clusters	1871	677	572	622	1871
Pseudo R^2	0.020	0.019	0.022	0.031	0.026

Notes: Standard errors are clustered at the subject level. * p<0.1, ** p<0.05, *** p<0.01.

Table C.7: Replication of Table 3.3 models with outliers dropped

Dependent variable: *Choice*	Number of displayed candidates			Full sample
	2	*4*	*8*	
Attractiveness	0.13**	0.13***	0.17***	0.13**
	(0.07)	(0.06)	(0.06)	(0.07)
Attractiveness * 4 worker treatment				-0.00
				(0.09)
Attractiveness * 8 worker treatment				0.04
				(0.09)
Female prop	0.45***	0.25***	0.47***	0.45***
	(0.11)	(0.10)	(0.10)	(0.11)
Female prop * 4 worker treatment				-0.20
				(0.15)
Female prop * 8 worker treatment				0.03
				(0.15)
Asian prop	0.13	0.33**	0.34**	0.13
	(0.16)	(0.15)	(0.16)	(0.16)
Asian prop * 4 worker treatment				0.20
				(0.22)
Asian prop * 8 worker treatment				0.21
				(0.23)
Black prop	-0.03	-0.16	-0.22	-0.03
	(0.14)	(0.13)	(0.14)	(0.14)
Black prop * 4 worker treatment				-0.13
				(0.19)
Black prop * 8 worker treatment				-0.19
				(0.20)
Latino prop	0.00	-0.01	-0.20	0.00
	(0.24)	(0.23)	(0.25)	(0.24)
Latino prop * 4 worker treatment				-0.01
				(0.33)
Latino prop * 8 worker treatment				-0.20
				(0.35)
N	1630	2352	3744	7726
Number of clusters	815	588	468	1871
Pseudo R^2	0.025	0.014	0.026	0.022

Notes: Standard errors are clustered at the subject level. * p<0.1, ** p<0.05, *** p<0.01.

C.3 Odds Ratio

Below are tables showing the odds ratios for each table presented in the main text. That is, if a coefficient in a table was estimated as *X*, the coefficient in the corresponding table shows e^X. These tables can be used to compute effect sizes by subtracting 1 from the relevant odds ratio.

Table C.8: Odds Ratios of the estimates from Table 3.2

Dependent variable: *Choice*	Full sample	Prior performance shown			
		none	*individual*	*subgroup*	*All*
Attractiveness	1.15*** (0.04)	1.28*** (0.07)	1.03 (0.06)	1.15*** (0.04)	1.27*** (0.07)
Attractiveness * Individual info treatment					0.81*** (0.07)
Attractiveness * Subgroup info treatment					0.92 (0.07)
Female prop	1.60*** (0.09)	1.38*** (0.13)	1.98*** (0.19)	1.52*** (0.15)	1.43*** (0.13)
Female prop * Individual info treatment					1.37** (0.18)
Female prop * Subgroup info treatment					1.06 (0.15)
Asian prop	1.23** (0.10)	1.33* (0.19)	0.95 (0.15)	1.43** (0.21)	1.35** (0.20)
Asian prop * Individual info treatment					0.69* (0.15)
Asian prop * Subgroup info treatment					1.07 (0.22)
Black prop	0.85** (0.06)	1.02 (0.13)	0.85 (0.11)	0.68*** (0.09)	1.03 (0.13)
Black prop * Individual info treatment					0.81 (0.14)
Black prop * Subgroup info treatment					0.66** (0.12)
Latino prop	0.91 (0.11)	1.28 (0.25)	0.82 (0.18)	0.69 (0.16)	1.23 (0.24)
Latino prop * Individual info treatment					0.66 (0.20)
Latino prop * Subgroup info treatment					0.56* (0.17)
N	9256	3320	2846	3090	9256
Number of clusters	2216	796	703	717	2216
Pseudo R^2	0.024	0.025	0.032	0.031	0.030

Notes: Robust standard errors are in parenthesis. * p<0.1, ** p<0.05, *** p<0.01.

Table C.9: Odds Ratios of the estimates from Table 3.3

Dependent variable: *Choice*	Number of displayed candidates			Full sample
	2	*4*	*8*	
Attractiveness	1.15*** (0.07)	1.14** (0.06)	1.16*** (0.07)	1.15** (0.07)
Attractiveness * 4 worker treatment				0.99 (0.08)
Attractiveness * 8 worker treatment				1.01 (0.08)
Female prop	1.73*** (0.18)	1.34*** (0.12)	1.86*** (0.18)	1.73*** (0.18)
Female prop * 4 worker treatment				0.77* (0.11)
Female prop * 8 worker treatment				1.07 (0.15)
Asian prop	1.10 (0.17)	1.42** (0.20)	1.16 (0.17)	1.10 (0.17)
Asian prop * 4 worker treatment				1.29 (0.27)
Asian prop * 8 worker treatment				1.06 (0.22)
Black prop	0.92 (0.12)	0.87 (0.11)	0.74** (0.09)	0.92 (0.12)
Black prop * 4 worker treatment				0.95 (0.17)
Black prop * 8 worker treatment				0.81 (0.14)
Latino prop	0.93 (0.21)	0.98 (0.19)	0.79 (0.17)	0.93 (0.21)
Latino prop * 4 worker treatment				1.06 (0.32)
Latino prop * 8 worker treatment				0.85 (0.27)
N	1912	2736	4608	9256
Number of clusters	956	684	576	2216
Pseudo R^2	0.035	0.017	0.031	0.027

Notes: Robust standard errors in parenthesis. * p<0.1, ** p<0.05, *** p<0.01.

Table C.10: Odds Ratios of the estimates from Table 3.4

Dependent variable: *Choice*	(1)	(2)
Attractiveness	1.08* (0.05)	1.08** (0.04)
Female prop	1.62*** (0.09)	1.71*** (0.11)
Female * Attractiveness	1.05 (0.06)	
Asian prop	1.22** (0.10)	1.09 (0.10)
Black prop	0.84** (0.06)	0.93 (0.08)
Latino prop	0.89 (0.11)	0.92 (0.13)
Prediction		2.91*** (0.13)
N	9256	9256
Number of clusters	2216	2216
Pseudo R^2	0.025	0.024

Notes: Robust standard errors in parenthesis. * p<0.1, ** p<0.05, *** p<0.01.

C.4 Details of Experimental Trials

The order of trials was randomized across participants. The order of candidates (i.e., workers) within each trial was also randomized. Also, whether information about past performance was displayed was randomized across participants (i.e., between subjects' experimental manipulation).

C.4.1 Two Workers Condition

Tables below show examples of experimental conditions in two, four and eight workers conditions with the past performance displayed at the individual level. In these tables, for example Male 5 means a Male who got 5 questions correct; Female 4 - a Female who got 4 questions correct. For example, in the second trial in Table C.11, participants had to choose between a male who got 2 questions correct, and a female that got 2 questions correct.

Table C.11: Example of past performance at individual level - two worker condition

Trial #	Candidate 1	Candidate 2
1	Male 0	Female 0
2	Male 2	Female 2
3	Male 4	Female 4
4	Male 5	Female 5
5	Male 1	Female 2
6	Male 2	Female 3
7	Male 1	Female 4
8	Male 2	Female 5
9	Male 2	Female 1
10	Male 3	Female 2
11	Male 4	Female 1
12	Male 5	Female 2

C.4.2 Four Workers Condition

Table C.12: Example of past performance at individual level - four worker condition

Trial #	Candidate 1	Candidate 2	Candidate 3	Candidate 4
1	Male 0	Female 0	Male 0	Female 0
2	Male 2	Female 2	Male 2	Female 2
3	Male 1	Female 4	Male 4	Female 1
4	Male 2	Female 2	Male 4	Female 4
5	Male 4	Female 4	Male 1	Female 1
6	Male 3	Female 1	Male 1	Female 3
7	Male 4	Female 5	Male 1	Female 1
8	Male 5	Female 4	Male 1	Female 1
9	Male 4	Female 1	Male 1	Female 4
10	Male 3	Female 4	Male 1	Female 1
11	Male 4	Female 3	Male 1	Female 1
12	Male 4	Female 2	Male 2	Female 4

C.4.3 Eight Workers Condition

Table C.13: Example of past performance at individual level - eight worker condition

Trial #	Candidate 1	Candidate 2	Candidate 3	Candidate 4	Candidate 5	Candidate 6	Candidate 7	Candidate 8
1	Male 3	Female 3	Male 3	Female 3	Male 3	Female 3	Male 3	Female 3
2	Male 4	Female 1	Male 0	Female 4	Male 1	Female 1	Male 1	Female 0
3	Male 0	Female 4	Male 1	Female 1	Male 4	Female 2	Male 1	Female 0
4	Male 5	Female 2	Male 2	Female 2	Male 1	Female 5	Male 2	Female 1
5	Male 0	Female 0	Male 5	Female 4	Male 1	Female 2	Male 2	Female 1
6	Male 0	Female 0	Male 4	Female 5	Male 1	Female 2	Male 2	Female 1
7	Male 1	Female 3	Male 0	Female 1	Male 5	Female 1	Male 0	Female 1
8	Male 0	Female 4	Male 0	Female 1	Male 3	Female 0	Male 1	Female 0
9	Male 0	Female 0	Male 3	Female 1	Male 0	Female 3	Male 1	Female 0
10	Male 1	Female 4	Male 0	Female 0	Male 4	Female 1	Male 1	Female 0
11	Male 5	Female 0	Male 0	Female 5	Male 1	Female 1	Male 2	Female 2
12	Male 1	Female 1	Male 4	Female 0	Male 1	Female 4	Male 1	Female 1

C.4.4 Information Provision Treatment

The tables in the condition where no information about the prior performance was displayed are identical to the above tables, except that prior performance at the individual level was not displayed.

The tables in the condition where only subgroup performance by gender was displayed are identical to the above tables, except that prior performance at the individual level was not displayed, and additionally, Figure 3.2 in the main text was displayed to participants at the start of the study, and they could click on a link to see the figure again at any point in the experiment if they wanted.

C.5 Recruiter Gender

This section replicates Tables 3.2, 3.3 and 3.4 disaggregated by gender of the recruiter. 57% of our recruiter subjects were male and remaining 43% were female.

C.5.1 Female Recruiter

Table C.14: Replication of Table 3.2 – Female recruiter

Dependent variable: *Choice*	Full sample	Prior performance shown			
		none	*individual*	*subgroup*	*All*
Attractiveness	0.07 (0.05)	0.16* (0.08)	-0.03 (0.11)	0.07 (0.09)	0.16* (0.08)
Attractiveness * Individual info treatment					-0.19 (0.13)
Attractiveness * Subgroup info treatment					-0.09 (0.12)
Female prop	0.44*** (0.09)	0.40*** (0.13)	1.07*** (0.18)	0.04 (0.15)	0.40*** (0.13)
Female prop * Individual info treatment					0.67** (0.22)
Female prop * Subgroup info treatment					-0.36* (0.20)
Asian prop	0.20 (0.13)	0.28 (0.20)	-0.11 (0.28)	0.34 (0.22)	0.28 (0.20)
Asian prop * Individual info treatment					-0.39 (0.35)
Asian prop * Subgroup info treatment					0.06 (0.30)
Black prop	-0.13 (0.11)	0.08 (0.18)	-0.15 (0.23)	-0.34* (0.19)	0.08 (0.18)
Black prop * Individual info treatment					-0.23 (0.29)
Black prop * Subgroup info treatment					-0.42 (0.26)
Latino prop	-0.00 (0.19)	0.25 (0.30)	-0.08 (0.39)	-0.10 (0.31)	0.25 (0.30)
Latino prop * Individual info treatment					-0.33 (0.49)
Latino prop * Subgroup info treatment					-0.36 (0.43)
N	3848	1550	932	1366	3848
Number of clusters	952	403	238	311	952
Pseudo R^2	0.018	0.023	0.071	0.012	0.031

Notes: Standard errors are clustered at the subject level. * p<0.1, ** p<0.05, *** p<0.01.

Table C.15: Replication of Table 3.3 – Female recruiter

Dependent variable: *Choice*	Number of displayed candidates			Full sample
	2	*4*	*8*	
Attractiveness	0.17* (0.09)	-0.11 (0.09)	0.13 (0.09)	0.17* (0.09)
Attractiveness * 4 worker treatment				-0.28** (0.13)
Attractiveness * 8 worker treatment				-0.04 (0.13)
Female prop	0.70*** (0.15)	0.28* (0.15)	0.36** (0.14)	0.70*** (0.15)
Female prop * 4 worker treatment				-0.42* (0.22)
Female prop * 8 worker treatment				-0.34* (0.21)
Asian prop	0.08 (0.23)	0.38 (0.23)	0.15 (0.22)	0.08 (0.23)
Asian prop * 4 worker treatment				0.30 (0.32)
Asian prop * 8 worker treatment				0.07 (0.32)
Black prop	-0.16 (0.19)	0.08 (0.21)	-0.34* (0.20)	-0.16 (0.19)
Black prop * 4 worker treatment				0.24 (0.28)
Black prop * 8 worker treatment				-0.19 (0.27)
Latino prop	-0.09 (0.32)	0.32 (0.33)	-0.15 (0.34)	-0.09 (0.32)
Latino prop * 4 worker treatment				0.40 (0.46)
Latino prop * 8 worker treatment				-0.05 (0.47)
N	920	1008	1920	3848
Number of clusters	460	252	240	952
Pseudo R^2	0.054	0.011	0.015	0.024

Notes: Standard errors are clustered at the subject level. * p<0.1, ** p<0.05, *** p<0.01.

Table C.16: Replication of Table 3.4 – Female recruiter

Dependent variable: *Choice*	(1)	(2)
Attractiveness	0.05	0.01
	(0.07)	(0.06)
Female prop	0.45***	0.48***
	(0.09)	(0.10)
Female * Attractiveness	0.01	
	(0.09)	
Asian prop	0.19	0.09
	(0.13)	(0.15)
Black prop	-0.13	0.09
	(0.11)	(0.13)
Latino prop	-0.00	0.21
	(0.19)	(0.23)
Prediction		1.31***
		(0.08)
N	3848	3848
Number of clusters	952	952
Pseudo R^2	0.018	0.298

Notes: Standard errors are clustered at the subject level. * p<0.1, ** p<0.05, *** p<0.01.

C.5.2 Male Recruiter

Table C.17: Replication of Table 3.2 – Male recruiter

Dependent variable: *Choice*	Full sample	Prior performance shown			
		none	*individual*	*subgroup*	*All*
Attractiveness	0.20*** (0.04)	0.32*** (0.08)	0.08 (0.07)	0.21*** (0.08)	0.32*** (0.08)
Attractiveness * Individual info treatment					-0.24** (0.10)
Attractiveness * Subgroup info treatment					-0.10 (0.11)
Female prop	0.52*** (0.07)	0.32** (0.13)	0.52*** (0.12)	0.73*** (0.14)	0.32** (0.13)
Female prop * Individual info treatment					0.20 (0.13)
Female prop * Subgroup info treatment					0.41** (0.19)
Asian prop	0.21* (0.11)	0.32 (0.21)	-0.07 (0.19)	0.37* (0.19)	0.32 (0.21)
Asian prop * Individual info treatment					-0.39 (0.28)
Asian prop * Subgroup info treatment					0.05 (0.28)
Black prop	-0.22** (0.10)	0.00 (0.18)	-0.24 (0.15)	-0.39*** (0.18)	0.00 (0.18)
Black prop * Individual info treatment					-0.24 (0.23)
Black prop * Subgroup info treatment					-0.39 (0.25)
Latino prop	-0.19 (0.17)	0.16 (0.27)	-0.21 (0.27)	-0.63* (0.34)	0.16 (0.27)
Latino prop * Individual info treatment					-0.38 (0.38)
Latino prop * Subgroup info treatment					-0.79* (0.43)
N	5384	1770	1890	1724	5384
Number of clusters	1252	393	453	406	1252
Pseudo R^2	0.033	0.032	0.021	0.064	0.038

Notes: Standard errors are clustered at the subject level. * p<0.1, ** p<0.05, *** p<0.01.

Table C.18: Replication of Table 3.3 – Male recruiter

Dependent variable: *Choice*	Number of displayed candidates			Full sample
	2	*4*	*8*	
Attractiveness	0.12 (0.09)	0.26*** (0.07)	0.17** (0.07)	0.12 (0.09)
Attractiveness * 4 worker treatment				0.14 (0.11)
Attractiveness * 8 worker treatment				0.04 (0.11)
Female prop	0.45*** (0.14)	0.29** (0.12)	0.82*** (0.13)	0.45*** (0.14)
Female prop * 4 worker treatment				-0.16 (0.18)
Female prop * 8 worker treatment				0.37* (0.19)
Asian prop	0.07 (0.21)	0.34* (0.18)	0.16 (0.19)	0.07 (0.21)
Asian prop * 4 worker treatment				0.27 (0.28)
Asian prop * 8 worker treatment				0.09 (0.28)
Black prop	-0.10 (0.18)	-0.26* (0.16)	-0.26 (0.16)	-0.10 (0.18)
Black prop * 4 worker treatment				-0.16 (0.24)
Black prop * 8 worker treatment				-0.16 (0.24)
Latino prop	-0.02 (0.32)	-0.21 (0.26)	-0.28 (0.28)	-0.02 (0.32)
Latino prop * 4 worker treatment				0.19 (0.41)
Latino prop * 8 worker treatment				-0.27 (0.43)
N	968	1728	2688	5384
Number of clusters	484	432	336	1252
Pseudo R^2	0.024	0.031	0.048	0.037

Notes: Standard errors are clustered at the subject level. * p<0.1, ** p<0.05, *** p<0.01.

Table C.19: Replication of Table 3.4 – Male recruiter

Dependent variable: *Choice*	(1)	(2)
Attractiveness	0.10* (0.06)	0.12*** (0.05)
Female prop	0.52*** (0.07)	0.57*** (0.08)
Female * Attractiveness	0.07 (0.08)	
Asian prop	0.19* (0.11)	0.09 (0129)
Black prop	-0.22** (0.10)	-0.17 (0.11)
Latino prop	-0.20 (0.17)	-0.25 (0.18)
Prediction		0.92*** (0.05)
N	5384	5384
Number of clusters	1252	1252
Pseudo R^2	0.033	0.208

Notes: Standard errors are clustered at the subject level. * p<0.1, ** p<0.05, *** p<0.01.

C.6 Information Provided to Employer Subjects

C.6.1 No Prior Performance Info Provided

Table C.20: Replication of Table 3.3 - No prior performance info provided

Dependent variable: *Choice*	Number of displayed candidates			Full sample
	2	*4*	*8*	
Attractiveness	0.21** (0.11)	0.14 (0.09)	0.35*** (0.10)	0.21** (0.11)
Attractiveness * 4 worker treatment				-0.07 (0.14)
Attractiveness * 8 worker treatment				0.14 (0.14)
Female prop	0.52*** (0.17)	-0.23 (0.15)	0.94*** (0.17)	0.52*** (0.17)
Female prop * 4 worker treatment				-0.75*** (0.23)
Female prop * 8 worker treatment				0.42* (0.24)
Asian prop	0.35 (0.27)	0.48** (0.24)	0.09 (0.25)	0.35 (0.27)
Asian prop * 4 worker treatment				0.13 (0.36)
Asian prop * 8 worker treatment				-0.26 (0.37)
Black prop	0.23 (0.22)	0.07 (0.21)	-0.18 (0.22)	0.23 (0.22)
Black prop * 4 worker treatment				-0.16 (0.30)
Black prop * 8 worker treatment				-0.40 (0.31)
Latino prop	0.46 (0.35)	0.06 (0.32)	0.24 (0.36)	0.46 (0.35)
Latino prop * 4 worker treatment				-0.39 (0.47)
Latino prop * 8 worker treatment				-0.21 (0.51)
N	680	1008	1632	3320
Number of clusters	340	252	204	796
Pseudo R^2	0.048	0.012	0.078	0.048

Notes: Standard errors are clustered at the subject level. * p<0.1, ** p<0.05, *** p<0.01.

C.6.2 Individual Performance Info Provided

Table C.21: Replication of Table 3.3 - Individual performance info provided

Dependent variable: *Choice*	Number of displayed candidates			Full sample
	2	*4*	*8*	
Attractiveness	0.03 (0.12)	0.13 (0.09)	-0.09 (0.11)	0.03 (0.12)
Attractiveness * 4 worker treatment				0.10 (0.15)
Attractiveness * 8 worker treatment				-0.12 (0.16)
Female prop	0.81*** (0.19)	-0.63*** (0.14)	0.64*** (0.19)	0.81*** (0.19)
Female prop * 4 worker treatment				-0.17 (0.24)
Female prop * 8 worker treatment				0.17 (0.27)
Asian prop	-0.11 (0.30)	0.32 (0.23)	-0.66* (0.19)	-0.11 (0.30)
Asian prop * 4 worker treatment				0.43 (0.38)
Asian prop * 8 worker treatment				0.56 (0.42)
Black prop	0.02 (0.25)	-0.17 (0.18)	-0.40* (0.24)	0.02 (0.25)
Black prop * 4 worker treatment				-0.19 (0.31)
Black prop * 8 worker treatment				-0.42 (0.35)
Latino prop	-0.25 (0.46)	-0.06 (0.33)	-0.45 (0.41)	-0.25 (0.46)
Latino prop * 4 worker treatment				0.20 (0.56)
Latino prop * 8 worker treatment				-0.20 (0.61)
N	542	1152	1152	2846
Number of clusters	271	288	144	703
Pseudo R^2	0.056	0.039	0.028	0.039

Notes: Standard errors are clustered at the subject level. * p<0.1, ** p<0.05, *** p<0.01.

C.6.3 Subgroup Performance Info Provided

Table C.22: Replication of Table 3.3 - Subgroup performance info provided

Dependent variable: *Choice*	Number of displayed candidates			Full sample
	2	*4*	*8*	
Attractiveness	0.18* (0.10)	0.14 (0.12)	0.16* (0.09)	0.18* (0.10)
Attractiveness * 4 worker treatment				-0.04 (0.16)
Attractiveness * 8 worker treatment				-0.02 (0.14)
Female prop	0.38** (0.17)	0.59** (0.22)	0.35** (0.15)	0.38** (0.17)
Female prop * 4 worker treatment				0.21 (0.28)
Female prop * 8 worker treatment				-0.03 (0.23)
Asian prop	0.03 (0.24)	0.26 (0.29)	0.67*** (0.22)	0.03 (0.24)
Asian prop * 4 worker treatment				0.23 (0.38)
Asian prop * 8 worker treatment				0.63* (0.32)
Black prop	-0.47** (0.21)	-0.41 (0.28)	-0.29 (0.20)	-0.47** (0.21)
Black prop * 4 worker treatment				0.07 (0.35)
Black prop * 8 worker treatment				0.18 (0.30)
Latino prop	-0.44 (0.38)	-0.05 (0.47)	-0.48 (0.37)	-0.44 (0.38)
Latino prop * 4 worker treatment				0.39 (0.61)
Latino prop * 8 worker treatment				-0.04 (0.54)
N	690	576	1824	3090
Number of clusters	345	144	228	717
Pseudo R^2	0.033	0.040	0.033	0.035

Notes: Standard errors are clustered at the subject level. * p<0.1, ** p<0.05, *** p<0.01.

C.7 Other Characteristics by Gender

This section reviews the results of "other characteristics" by gender of applicants. As the results in the two tables below indicate, the impact of *Attractiveness* relates to female applicants, with beauty irrelevant for male applicants. Similarly, *Masculine* and *Feminine* impact is also driven by female applicants. On the other hand, *Asian* and *Dominant* effects are mainly driven by male applicants. The impact of the rest of the characteristics (*Black*, *Latino*, *Angry*, *Happy*, *Trustworthy* and *Threatening*) are not driven by any particular gender.

Table C.23: Estimates of Discrete Choice Model – Other characteristics: Female applicants

Dependent variable: *Choice*	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Attractiveness	0.21*** (0.06)	0.22*** (0.06)	0.13** (0.06)	0.07 (0.07)	0.22*** (0.05)	0.17*** (0.06)	0.20*** (0.06)
Asian prop	0.21 (0.14)	0.22 (0.14)	0.17 (0.14)	0.20 (0.14)	0.18 (0.14)	0.17 (0.14)	0.18 (0.14)
Black prop	-0.18* (0.11)	-0.18 (0.11)	-0.18 (0.11)	-0.18 (0.11)	-0.16 (0.11)	-0.20* (0.11)	-0.18* (0.11)
Latino prop	-0.09 (0.19)	-0.09 (0.19)	-0.17 (0.19)	-0.11 (0.19)	-0.06 (0.19)	-0.07 (0.19)	-0.09 (0.19)
Angry	-0.04 (0.07)						
Happy		-0.04 (0.04)					
Masculine			-0.13** (0.06)				
Feminine				0.18** (0.07)			
Dominant					-0.13*** (0.05)		
Trustworthy						-0.13 (0.09)	
Threatening							-0.14*** (0.05)
N	2896	2896	2896	2896	2896	2896	2896
Number of clusters	977	977	977	977	977	977	977
Pseudo R^2	0.018	0.018	0.021	0.022	0.019	0.022	0.018

Notes: Standard errors are clustered at the subject level. * p<0.1, ** p<0.05, *** p<0.01. Various characteristics from the Chicago Face Database are added to the main specification in this table.

Table C.24: Estimates of Discrete Choice Model – Other characteristics: Male applicants

Dependent variable: *Choice*	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Attractiveness	0.05 (0.08)	0.09 (0.09)	0.11 (0.08)	0.08 (0.08)	0.11 (0.08)	0.05 (0.10)	0.05 (0.08)
Asian prop	0.28* (0.16)	0.30* (0.16)	0.27* (0.16)	0.31* (0.16)	0.23 (0.17)	0.25 (0.17)	0.23 (0.17)
Black prop	-0.15 (0.15)	-0.15 (0.15)	-0.11 (0.15)	-0.14 (0.15)	-0.09 (0.15)	-0.18 (0.15)	-0.15 (0.15)
Latino prop	0.00 (0.25)	-0.00 (0.25)	0.03 (0.25)	-0.00 (0.25)	-0.10 (0.26)	0.02 (0.25)	0.01 (0.25)
Angry	-0.13 (0.09)						
Happy		0.00 (0.09)					
Masculine			-0.10 (0.07)				
Feminine				0.09 (0.07)			
Dominant					-0.19*** (0.09)		
Trustworthy						-0.14 (0.18)	
Threatening							-0.16 (0.10)
N	1787	1787	1787	1787	1787	1787	1787
Number of clusters	643	643	643	643	643	643	643
Pseudo R^2	0.008	0.006	0.008	0.007	0.009	0.006	0.008

Notes: Standard errors are clustered at the subject level. * p<0.1, ** p<0.05, *** p<0.01. Various characteristics from the Chicago Face Database are added to the main specification in this table.

C.8 Other Characteristics by Race

In this section, we review the results of "other characteristics" by the race of applicants. The tables below show that the strong positive impact of *Attractiveness* is driven by *White* applicants, with beauty not a significant factor for other races. We also see that the impact of *Female* is largely homogeneous by the race of applicants: for all races, *Females* have a large positive (and in most cases significant) impact on hiring probability. As for "other characteristics", *White* applicants are the main driver of the impact of *Masculine* and *Feminine*, while *Asians* drive a negative impact for *Angry* faces. The *Dominant* impact is driven by *Asian* and *Black* applicants, with the rest of the characteristics (*Happy*, *Trustworthy* and *Threatening*) are not driven by any particular race.

Table C.25: Estimates of Discrete Choice Model – Other characteristics: Asian applicants

Dependent variable: *Choice*	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Attractiveness	-0.00 (0.14)	0.04 (0.16)	0.10 (0.14)	-0.11 (0.18)	0.08 (0.13)	0.03 (0.16)	0.05 (0.14)
Female prop	0.77*** (0.21)	0.74*** (0.21)	0.79 (0.68)	-0.45 (0.74)	0.59*** (0.21)	0.71*** (0.21)	0.70*** (0.21)
Angry	-0.27* (0.15)						
Happy		0.08 (0.15)					
Masculine			0.04 (0.27)				
Feminine				0.42 (0.26)			
Dominant					-0.44** (0.19)		
Trustworthy						0.18 (0.28)	
Threatening							-0.17 (0.19)
N	622	622	622	622	622	622	622
Number of clusters	253	253	253	253	253	253	253
Pseudo R^2	0.061	0.054	0.053	0.059	0.065	0.054	0.055

Notes: Standard errors are clustered at the subject level. * p<0.1, ** p<0.05, *** p<0.01. Various characteristics from the Chicago Face Database are added to the main specification in this table.

Table C.26: Estimates of Discrete Choice Model – Other characteristics: Black applicants

Dependent variable: *Choice*	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Attractiveness	0.09 (0.12)	0.13 (0.12)	0.13 (0.14)	-0.01 (0.22)	0.14 (0.10)	0.23* (0.12)	0.11 (0.12)
Female prop	0.71*** (0.16)	0.70*** (0.16)	0.43 (0.70)	-0.04 (0.81)	0.49*** (0.18)	0.70*** (0.16)	0.66*** (0.16)
Angry	-0.17 (0.14)						
Happy		0.09 (0.13)					
Masculine			-0.10 (0.24)				
Feminine				0.27 (0.28)			
Dominant					-0.37** (0.16)		
Trustworthy						-0.20 (0.27)	
Threatening							-0.17 (0.19)
N	829	829	829	829	829	829	829
Number of clusters	322	322	322	322	322	322	322
Pseudo R^2	0.051	0.049	0.049	0.050	0.056	0.049	0.050

Notes: Standard errors are clustered at the subject level. * $p<0.1$, ** $p<0.05$, *** $p<0.01$. Various characteristics from the Chicago Face Database are added to the main specification in this table.

Table C.27: Estimates of Discrete Choice Model – Other characteristics: Latino applicants

Dependent variable: *Choice*	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Attractiveness	-0.16 (0.17)	-0.09 (0.19)	-0.17 (0.17)	-0.07 (0.18)	-0.15 (0.17)	-0.15 (0.19)	-0.17 (0.17)
Female prop	0.49*** (0.22)	0.41* (0.23)	0.72 (0.54)	1.32 (0.82)	0.47** (0.22)	0.48** (0.21)	0.47** (0.21)
Angry	-0.03 (0.19)						
Happy		-0.15 (0.18)					
Masculine			0.08 (0.17)				
Feminine				-0.30 (0.28)			
Dominant					-0.01 (0.17)		
Trustworthy						-0.01 (0.37)	
Threatening							-0.14 (0.20)
N	457	457	457	457	457	457	457
Number of clusters	191	191	191	191	191	191	191
Pseudo R^2	0.017	0.019	0.018	0.021	0.017	0.017	0.019

Notes: Standard errors are clustered at the subject level. * p<0.1, ** p<0.05, *** p<0.01. Various characteristics from the Chicago Face Database are added to the main specification in this table.

Table C.28: Estimates of Discrete Choice Model – Other characteristics: White applicants

Dependent variable: *Choice*	(1)	(2)	(3)	(4)	5	6	7
Attractiveness	-0.24*** (0.08)	-0.23** (0.09)	-0.17** (0.08)	-0.07 (0.11)	-0.23** (0.08)	-0.28*** (0.09)	-0.22*** (0.08)
Female prop	0.42*** (0.13)	0.42*** (0.13)	-0.34 (0.37)	-0.49 (0.45)	0.42*** (0.14)	0.45*** (0.13)	0.42*** (0.13)
Angry	0.04 (0.10)						
Happy		0.00 (0.12)					
Masculine			-0.31** (0.14)				
Feminine				0.34** (0.16)			
Dominant					-0.00 (0.11)		
Trustworthy						-0.21 (0.22)	
Threatening							-0.04 (0.12)
N	1133	1133	1133	1133	1133	1133	1133
Number of clusters	421	421	421	421	421	421	421
Pseudo R^2	0.033	0.033	0.040	0.039	0.033	0.034	0.033

Notes: Standard errors are clustered at the subject level. * p<0.1, ** p<0.05, *** p<0.01. Various characteristics from the Chicago Face Database are added to the main specification in this table.