Blueprint Labs

Discussion Paper #2021.07

Systemic Discrimination Among Large U.S. Employers

Patrick M. Kline
Evan K. Rose
Christopher R. Walters

July 2021

MIT Department of Economics
77 Massachusetts Avenue, Bldg. E53-390
Cambridge, MA 02139

National Bureau of Economic Research
1050 Massachusetts Avenue, 3rd Floor
Cambridge, MA 02138

NBER WORKING PAPER SERIES

Blueprint Labs Working Paper #2021.07

SYSTEMIC DISCRIMINATION AMONG LARGE U.S. EMPLOYERS

Patrick M. Kline
Evan K. Rose
Christopher R. Walters

Working Paper 29053
http://www.nber.org/papers/w29053

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2021

We thank Sendhil Mullainathan, Sapna Raj, and Jenny Yang for helpful conversations, and seminar participants at UC Berkeley, Chicago Booth, the University of Chicago HCEO Working Group, CEPR, IZA, and CESifo for useful comments. We also thank the J-PAL North America Social Policy Research Initiative for generous funding support. A pre-analysis plan for this project can be found under AEA RCT registry number AEARCTR-0004739. Outstanding research assistance was provided by Jake Anderson, Hadar Avivi, Ross Chu, Brian Collica, Nicole Gandre, and Ben Scuderi. We are grateful to our determined team of undergraduate volunteers, without whom this project would not have been possible: May Adberg, Jason Chen, Stephanie Cong, Simon Duabis, Daniel Dychala, Samuel Gao, Alexandra Groscost, Victoria Haworth, Camille Hillion, Ben Keltner, Mary Kruberg, Jiaxin Lei, Carol Lee, Collin Lu, Oliver McNeil, Riley Odom, Sarah Phung, Eric Phillips, Stephanie Ross, Marcus Sander, Pat Tagari, Quinghuai Tan, Lydia Wen, Zijun Xu, Xilin Ying, Andy Zhong, Leila Zhua, Yingjia Zhang, and Yiran Zi. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

Systemic Discrimination Among Large U.S. Employers
Patrick M. Kline, Evan K. Rose, and Christopher R. Walters
NBER Working Paper No. 29053
July 2021
JEL No. C11,C9,C93,J7,J71,J78,K31,K42

ABSTRACT

We study the results of a massive nationwide correspondence experiment sending more than 83,000 fictitious applications with randomized characteristics to geographically dispersed jobs posted by 108 of the largest U.S. employers. Distinctively Black names reduce the probability of employer contact by 2.1 percentage points relative to distinctively white names. The magnitude of this racial gap in contact rates differs substantially across firms, exhibiting a between-company standard deviation of 1.9 percentage points. Despite an insignificant average gap in contact rates between male and female applicants, we find a between-company standard deviation in gender contact gaps of 2.7 percentage points, revealing that some firms favor male applicants while others favor women. Company-specific racial contact gaps are temporally and spatially persistent, and negatively correlated with firm profitability, federal contractor status, and a measure of recruiting centralization. Discrimination exhibits little geographical dispersion, but two digit industry explains roughly half of the cross-firm variation in both racial and gender contact gaps. Contact gaps are highly concentrated in particular companies, with firms in the top quintile of racial discrimination responsible for nearly half of lost contacts to Black applicants in the experiment. Controlling false discovery rates to the 5% level, 23 individual companies are found to discriminate against Black applicants. Our findings establish that systemic illegal discrimination is concentrated among a select set of large employers, many of which can be identified with high confidence using large scale inference methods.

Patrick M. Kline
Department of Economics
University of California, Berkeley
530 Evans Hall #3880
Berkeley, CA 94720
and NBER
pkline@econ.berkeley.edu

Evan K. Rose
Becker Friedman Institute
University of Chicago
Chicago, IL
ekrose@gmail.com

Christopher R. Walters
Department of Economics
University of California, Berkeley
530 Evans Hall #3880
Berkeley, CA 94720-3880
and NBER
crwalters@econ.berkeley.edu

1 Introduction

Employment discrimination is a stubbornly persistent social problem. Title VII of the Civil Right Act of 1964 forbids employment discrimination on the basis of race, sex, color, religion, and national origin. A large social science literature uses resume correspondence experiments to document that these protected characteristics influence employer treatment of job applications (Bertrand and Duflo, 2017; Baert, 2018; Quillian et al., 2017), with some studies finding that this disparate treatment predicts later hiring decisions (Quillian, Lee and Oliver, 2020). Less is known, however, about the extent to which discrimination varies across employers. Is the U.S. labor market characterized by a small faction of severe discriminators adrift in an ocean of unbiased firms, or do most companies exhibit roughly equivalent biases?

The answer to this question has a host of important ramifications. As emphasized by Becker (1957), if discrimination is confined to a small minority of firms, workers may be able to avoid prejudice by sorting to non-discriminatory firms. The organizational structure of discrimination is also of great interest to the U.S. Equal Employment Opportunity Commission (EEOC), which is charged with preventing and remedying unlawful employment discrimination. In fiscal year 2020, the EEOC launched 558 investigations into systemic discrimination, defined as "a pattern or practice, policy and/or class cases where the discrimination has a broad impact on an industry, profession, company or geographic location" (U.S. EEOC, 2006*b*). If discrimination is concentrated among a small minority of firms, it may be possible to identify those firms with sufficient confidence to enable better decisions by the EEOC and other regulatory bodies. Evidence of discrimination could also potentially be shared directly with firms or made public, allowing companies to preemptively reform their practices.

This paper reports the results of a massive nationwide correspondence experiment designed to measure patterns of discrimination by large U.S. companies. The experiment sent over 83,000 job applications to job vacancies posted by 108 Fortune 500 employers. For each company, we sought to sample 125 entry-level jobs in distinct U.S. counties. Each sampled job received eight fictitious applications, with phone lines and email addresses monitored for contact. By sampling a large number of geographically distinct jobs from each company, we are able to average out idiosyncrasies associated with particular geographic areas, establishments, or hiring managers, revealing organization-wide patterns.

Following a large social science literature (Bertrand and Duflo, 2017; Baert, 2018), the experiment manipulated employer perceptions of race by randomly assigning racially distinctive names to job applications. Each job received four pairs of applications, with one member of each pair assigned a distinctively Black name and the other a distinctively white name. We also randomly varied signals of applicant sex, age, sexual orientation, gender identity, and political leaning. To filter out the potential effects of fluctuations in aggregate labor market conditions, jobs were sampled in five waves, with firms sampled in random order in each wave.

We find that distinctively Black names reduce the likelihood of employer contact relative to distinctively white names by 2.1 percentage points, an effect equal to 9% of the Black mean contact rate. This gap in contact rates varies substantially across the firms in our study. After adjusting for sampling error, the cross-firm standard deviation of racial contact gaps is 1.9 percentage points, only slightly below the mean contact penalty for Black names. Racial discrimination is heterogeneous but one-sided: we cannot reject the null hypothesis that all 108 firms in our experiment weakly favor white applicants. An application of Efron's (2016) empirical Bayes (EB) deconvolution estimator reveals that the distribution of racial contact gaps is highly skewed, with most firms exhibiting mild discrimination against Black applicants but a few exhibiting very large biases. We estimate that the top quintile of discriminating firms are responsible for 46% of the lost contacts to Black applicants in our experiment.

Companies also vary enormously in their treatment of applicant gender. On average, male and female applicants are equally likely to be contacted, but the standard deviation of gender contact gaps across companies is 2.7 percentage points, with a distribution that is roughly symmetric about zero. This "bi-directional" discrimination result accords with the findings of Kline and Walters (2021), who conclude, using different methods, that some jobs sampled in a correspondence experiment of Mexican employers (Arceo-Gomez and Campos-Vazquez, 2014) discriminated against women, while others discriminated against men. Our analysis shows that a similar pattern holds across large U.S. employers with geographically isolated establishments. Like racial discrimination, gender discrimination is highly concentrated in particular firms, with the top quintile of discriminating firms responsible for 57% of contacts lost to gender discrimination.

While our main focus is on race and gender, we also assess the extent of discrimination on several other dimensions. We find a modest contact penalty of 0.6 percentage points for applicants listing high school graduation dates implying an age over 40. This gap also varies across employers, with a cross-firm standard deviation of 1.1 percentage points. In contrast to race, gender, and age, we find no significant penalty for membership in a Lesbian, Gay, Bisexual, Transgender, or Queer (LGBTQ) club or evidence of heterogeneity in that penalty across firms. Likewise, we find insignificant effects of listing gender-neutral pronouns next to an applicant's name, though estimates for LGBTQ clubs and gender-neutral pronouns are less precise than estimates for race, gender, and age.

Having shown that discrimination varies substantially across employers, we assess whether this variation is predicted by geography, industry, and other establishment and firm characteristics. Surprisingly, geographic variation in race, gender, and age discrimination is extremely muted. We cannot reject the null hypothesis that mean contact gaps for gender and age are equal across all 50 states, and we find only marginally significant evidence against this null for racial contact gaps. In contrast, we estimate that 2-digit Standard Industrial Classification (SIC) codes explain roughly half of firm-level variation in contact gaps for both race and gender. Race and gender contact gaps also vary significantly by job title but this variation is indistinguishable from noise conditional on firm fixed effects. Contact gaps exhibit limited variation across third party intermediaries that power firms' hiring websites.

Both racial and gender contact gaps are estimated to be larger at jobs requiring social interaction. Racial contact gaps are larger at establishments in geographic areas with more prejudice as measured by Implicit Association Tests (IATs) and internet searches for racial epithets. In line with the predictions of Becker (1957), racial contact gaps are smaller at more profitable firms. Contact gaps also tend to be smaller among federal contractors. Finally, we find that firms with more centralized points of contact (i.e., callbacks originating from the same phone numbers) have much smaller contact gaps, suggesting that human resources practices may be an important mediator of discrimination.

The finding of significant employer heterogeneity in discrimination motivates an investigation of which particular organizations are likely violating the Civil Rights Act. Following Efron (2016), we use non-parametric estimates of the cross-firm distribution of discrimination to form EB posterior mean estimates of the contact gap at each individual firm. Forty-six of the 108 firms in our sample have posterior mean racial contact gaps in excess of 2 percentage points. These firms are disproportionately clustered in the auto services and sales sector and certain forms of retail. We find large posterior mean contact gaps favoring women at apparel stores but favoring men in the wholesale durable sector.

While posterior means provide best predictions of the extent of discrimination at each firm, it is also of interest to provide an assessment of which companies are likely to be discriminating at all. Applying large-scale multiple testing techniques introduced by Storey (2002, 2003), we find that 23 of the firms in our study discriminate against Black applicants with at least 95% posterior certainty (i.e., controlling false discovery rates to no more than 5%). This implies that at least 22 of these 23 individual firms should be expected to have non-zero racial contact gaps. These discriminating firms are over-represented in the auto sector, in general merchandising, and among eating and drinking establishments. In contrast, we find only one firm that can be reliably labeled as discriminating against men, and are unable to reliably identify any firms that discriminate against women while limiting false discovery rates to 5%. Our sharper detection power for racial discrimination stems from the fact that a larger share of firms in the population are estimated to discriminate based upon race than upon gender, increasing the prior probability of discrimination used to draw inferences about the conduct of individual firms. The single firm identified as discriminating against men is an apparel retailer that also discriminates against Black applicants with high posterior certainty.

In principle, firm-wide contact gaps may be driven by a small share of heavily biased jobs. We develop a simple lower bound on the prevalence of job-level discrimination based on split-sample estimates of the job-level variance of contact gaps. At least 7% of all jobs in our experiment discriminate against distinctively Black names. Among the 23 firms we can reliably label as engaged in racial discrimination, at least 20% of the jobs discriminate against Black names. At the modal firm in this group, this bound implies racial discrimination took place in at least 25 distinct U.S. counties out of the 125 sampled in our experiment, indicating a widespread national pattern.

We conclude with an economic analysis of optimal auditing strategies meant to mimic the objectives and constraints of regulatory authorities such as the EEOC. A hypothetical auditor seeks to investigate firms with large racial contact gaps. Informational constraints limit the expected yield on audits relative to the first-best investigation rule. We show that auditing strategies controlling the false discovery rate can be justified by a scenario in which the auditor seeks to avoid investigations of non-discriminators and faces ambiguity regarding the share of discriminatory firms in the population. In practice, we find that making decisions based upon false discovery rates rather than posterior means yields little reduction in the expected yield on investigations. The set of 23 firms that we can reliably classify as discriminating against Black names are estimated to account for 40% of lost contacts to Black applicants in our experiment. Our findings demonstrate that it is possible to identify individual firms responsible for a substantial share of racial discrimination while maintaining a tight limit on the expected number of false positives encountered.

The rest of the paper is organized as follows. Section 2 provides background on employment discrimination and the law. Section 3 details the experimental design, Section 4 describes the data, and Section 5 reports basic experimental impacts. Section 6 documents variation in discrimination across firms, while Section 7 examines variation across other groupings of jobs. Section 8 investigates relationships between discrimination and observed employer characteristics. Section 9 reports estimates of the full distribution of discrimination across firms. Section 10 uses this distribution to construct posterior estimates for individual firms and assesses the conclusions that can be drawn about discrimination by specific employers. Section 11 considers the consequences of our findings for regulatory auditing decisions. Finally, Section 12 concludes with a discussion of implications for anti-discrimination policy and directions for future research.

2 Policy Background

Much of the economics literature has focused on separating the contributions of taste-based and statistical discrimination to observed disparities, an exercise that requires inferring the extent to which employer conduct is motivated by beliefs regarding the productivity of different groups of workers (Becker, 1957, 1993; Aigner and Cain, 1977; Charles and Guryan, 2008; Bohren et al., 2019). Recent empirical and methodological work looks at group differences in the treatment of equally-productive individuals in bail decisions, motor vehicle searches, probation revocations, and other settings (Arnold, Dobbie and Yang, 2018; Arnold, Dobbie and Hull, 2020; Canay, Mogstad and Mountjoy, 2020; Rose, 2020; Feigenberg and Miller, 2021; Hull, 2021). In the employment context, it is widely understood that both taste-based and statistical discrimination typically involve disparate treatment of individuals according to legally-protected characteristics, which is prohibited by the Civil Rights Act.[1]

This paper is concerned with measuring such illegal disparate treatment, however motivated. The correspondence experiment we study was designed to manipulate employer perceptions of protected characteristics. Although the legal standing of organizations eliciting evidence of discrimination via "testing" remains unresolved (U.S. EEOC, 1996), an employer whose decision to contact a job applicant is influenced by the applicant's perceived race or sex has nonetheless engaged in disparate treatment and nominally violated the provisions of the Civil Right Act.[2]

Our use of the word "systemic" is motivated by the legal definition of this term as a "pattern or practice" of discrimination (U.S. EEOC, 2006*b*). The EEOC's systemic cases may concern either patterns of disparate treatment on protected characteristics or practices that target non-protected characteristics but nonetheless have disparate impacts on protected groups.[3] Key to either sort of case is evidence that the pattern or practice is widespread, affecting a company's hiring behavior at multiple locations. While our analysis will not reveal the specific polices or practices giving rise to systemic discrimination, we will be able to assess whether a nationwide pattern of disparate treatment is present at particular companies. This information may be of use both to the EEOC and to local organizations interested in promoting fair hiring practices.[4] Evidence of patterns of dis-

[1] Federal guidelines clearly state that "an employer may not base hiring decisions on stereotypes and assumptions about a person's race, color, religion, sex (including gender identity, sexual orientation, and pregnancy), national origin, age (40 or older), disability or genetic information" (see `https://www.eeoc.gov/prohibited-employment-policiespractices`).

[2] See U.S. Equal Employment Opportunity Commission v. Target Corp 460 F.3d 946 (7th Cir. 2006), Onwuachi-Willig and Barnes (2005), and footnote 27 of Fryer Jr and Levitt (2004) for discussion of the legal ramifications of handling fictitious applications based upon perceptions regarding race. In cases where no aggrieved individual has claimed standing, the EEOC can file a *commissioner's charge* alleging Title VII violations or launch a *directed investigation* into violations of either the Age Discrimination in Employment Act or the Equal Pay Act. In fiscal years 2016-2019, the EEOC averaged 13 commissioner's charges and 138 directed investigations per year.

[3] For instance, in 2019, the EEOC brought a systemic lawsuit against Schuster trucking for subjecting job applicants to a physical abilities test that was alleged to have a disparate impact on women (U.S. EEOC, 2019). However another 2019 case, against Sactacular Holdings LLC, an adult retail chain, alleged disparate treatment after a male job applicant was told by employees at two separate stores that the company does not consider men for sales associate positions (U.S. EEOC, 2020).

[4] For example, the New York City Commission on Human Rights has a mandate to test for discrimination in housing and labor markets and has assisted in the staging of matched-pairs audits of bias by landlords (Fang, Guess and Humphreys, 2019) and employers (Pager, Bonikowski and Western, 2009).

parate treatment in hiring by federal contractors is especially pertinent to the Office of Federal Contract Compliance Programs (OFCCP), which has broad discretion to audit contractors for compliance with executive orders prohibiting employment discrimination.

In deciding whether to launch investigations or audits, federal agencies often rely on analyses of employment data. For instance, the "inexorable zero" standard of Justice O'Connor, which refers to the complete absence of a group from a company's employees, has been taken as an indicator of discrimination, despite the difficulties of ascertaining whether qualified applicants were actually passed over by the firm (Huang, 2004).[5] In contrast, the correspondence experiment we study directly manipulated employer perceptions, permitting inferences to be drawn regarding average causal effects of protected characteristics on employer conduct. Examining effects across a large set of establishments tests for a systemic pattern of discrimination. While such patterns may be driven by official hiring practices, they may also reflect implicit biases on the part of employees with hiring authority. In either case, documentation of such nationwide patterns can aid efforts to ensure compliance with the law.

3 Experimental Design

Our study aims to measure the distribution of discrimination across the largest employers in the U.S. Figure 1 summarizes the sampling frame for the experiment. We began with the Fortune 500, splitting holding companies into brands with separate proprietary hiring websites. Data from InfoGroup and Burning Glass were used to determine the geographic distribution of establishments and vacancies, and each company's hiring portal was investigated for compatibility with our auditing methods. We determined that 108 companies (i.e., separate brands with distinct hiring websites and systems) had both enough geographic variation and routinely posted enough entry-level jobs on an easily-accessible portal to satisfy our sampling criteria.

We sampled 125 entry-level job vacancies for each employer, each corresponding to an establishment in a different U.S. county. Sampling was organized in a series of 5 waves, with a target of 25 jobs sampled for each firm in each wave. As shown in Figure 1, 72 of the 108 firms were sampled in all waves; some firms were excluded from the first wave due to an interruption caused by the COVID-19 pandemic, while others were excluded in later waves due to new technological barriers in their job portals. We randomly ordered

[5] The EEOC compliance manual references this standard in its guidelines for evaluating systemic discrimination: "a pattern or practice would be established if, despite the fact that Blacks made up 20 percent of a company's applicants for manufacturing jobs and 22 percent of the available manufacturing workers, not one of the 87 jobs filled during a six year period went to a Black applicant" (U.S. EEOC, 2006*a*). As the Supreme Court notes in Teamsters v. United States (1977) "the proof of the pattern or practice supports an inference that any particular employment decision, during the period in which the discriminatory policy was in force, was made in pursuit of that policy."

firms at the beginning of each wave and moved sequentially through the list, sampling the most recent job posting in a new county for each firm and randomizing ties. Each sampled job received 8 job applications with randomized characteristics. This sampling protocol yields a sample size for each employer of 1,000 applications, spread across the 125 jobs, for a total target of approximately 100,000 applications.

Applications were sent to each job in pairs. To minimize the chances of detection by employers, we allowed a gap of 1-2 days between consecutive pairs.[6] Though some vacancies closed while applications were still being sent, 87% of sampled jobs received the full 8 applications and 99% of jobs received at least two. As a result of vacancy closures and the exclusion of some firms from some waves, our final sample size amounted to roughly 84,000 applications.

As in many previous experiments measuring discrimination (Bertrand and Duflo, 2017), we signal race using racially-distinctive names. Our database of distinctive first names starts with those used by Bertrand and Mullainathan (2004), who used 9 unique names for each race and gender group, and supplements this list with 10 additional names per group from a database of speeding tickets issued in North Carolina between 2006 and 2018. We classified a name as racially distinctive if more than 90 percent of individuals with that name are of a particular race, and selected the most common distinctive Black and white names for those born between 1974 and 1979. We assembled distinctive last names from the 2010 U.S. Census, selecting names with high race-specific shares among those that occur at least 10,000 times nationally.[7] Together with our database of first names, this list generates about 500 unique full names for each race and gender category. One application within each pair is randomly assigned a distinctively white name while the other is randomly assigned a distinctively Black name. We draw names without replacement to ensure that no two applications to the same firm shared a name.

Our experiment also randomly assigns other legally-protected applicant characteristics. Sex is conveyed by applicant names. Fifty-percent of our names are distinctively female, and the rest are distinctively male. Assignment of sex is not stratified; therefore, each job receives between zero and eight female applications. Applicants are randomly assigned a date of birth implying an age between 22 and 58 years old, with ages uniformly distributed over this range. Because the Age Discrimination Act of 1967 prohibits discrimination against individuals aged 40 or older, we focus on differences between applicants over and under 40.

In Bostock v. Clayton County, Georgia (2020), the U.S. Supreme Court ruled that

[6] Pairs were sent every other day during wave 1, when most applications were submitted by human research assistants, to manage workloads. Beginning in wave 2, when the majority of applications were submitted automatically by software we developed, one pair was sent per day. Some pairs were occasionally sent with longer time lags due to workload or technological constraints, but overall 94% of applications were sent within 8 days of the first.

[7] All names used are presented in Appendix B along with additional details on experimental design.

discrimination based upon sexual orientation or gender identity violates Title VII of the Civil Rights Act. We began measuring discrimination on these dimensions starting in wave 2 of the experiment. Sexual orientation is conveyed by randomly assigning 10% of applicants to list LGBTQ high school clubs on their resumes. To distinguish between sexual orientation and general effects of clubs we randomly assign an additional 10% of applicants to be members of political or academic clubs. We convey gender identity by randomly assigning pronouns to 10% of resumes. Half of resumes with pronouns are assigned gender-typical pronouns (he/him for applicants with male names, she/her for applicants with female names), and the other half receive gender-neutral pronouns (they/them). Pronouns are listed on applicants' PDF resumes below their names.

Each fictitious applicant receives a large set of additional characteristics. All applicants graduated from high school in the year of their 18th birthday, with school names drawn randomly from a set of public high schools near the target job. Half of applicants receive associate degrees. Work histories consist of two or three jobs with nearby employers providing relevant experience. For example, retail job applicants have employment experience at local restaurants and retailers. In addition to populating fields in the employer's online job portal, we also upload a formatted PDF resume where possible, with resume templates and formatting drawn from a database of possible layouts. Some example resumes are provided in Appendix Figure A1. For employers requiring personality tests or other assessments, we pre-populate all answers to the assessments and randomly assign responses subject to the constraint that the applicant must pass the assessment. Random assignment of all supplementary characteristics takes place automatically, and these characteristics are independent of legally protected attributes and each other.

Our primary outcome is whether an employer attempted to contact the fictitious applicant. Phone numbers and e-mail addresses assigned to the fictitious applicants were monitored to determine when employers reached out for an interview. Contact information was assigned to ensure that no two applicants to the same firm shared an e-mail address or phone number. Our analysis focuses on whether the employer attempted to contact an applicant by any method within 30 days of applying. We also report results for other follow-up windows and specific contact types. Further details on the experimental design are available in our registered pre-analysis plan and in Appendix B.[8]

4 Summary Statistics

Table 1 provides summary statistics on two analysis samples. The baseline sample includes all 108 firms that were included in at least one wave. As a robustness exercise, we also consider a second sample that restricts to the 72 firms sampled in all waves of the

[8] The pre-analysis plan is stored in the AEA RCT registry with number AEARCTR-0004739. The full pre-analysis plan will not be available for public download until the conclusion of the study.

experiment.

In both samples, roughly half of the applications were assigned distinctively Black names. The slight discrepancy between white and Black sample sizes arises because job vacancies were occasionally taken offline before the second application of a race-balanced pair could be submitted. As expected, other resume characteristics are balanced across Black and white applications. About half of applications in each group are female. Slightly more than half of applications have high school graduation dates implying ages over 40, a consequence of the fact that the set of applicant birth years was not updated between waves 1 and 2. In subsequent waves we updated birth years to maintain a mean age of 40. By chance, white resumes are slightly less likely than Black resumes to list an associate degree.

On average, roughly 24% of applications were contacted by firms within 30 days. Most of these contact attempts arrived within 14 days. While the most common form of contact was voicemail, a substantial minority of applications were contacted via email or text message. In what follows we pool these forms of contact together and focus on effects of protected characteristics on the probability of any contact.

5 Average Contact Gaps

Employers are significantly less likely to contact applicants with distinctively Black names. The bottom panel of Table 1 reveals that the contact rate in the 30 days following an application is 2.1 percentage points (9%) higher for white applications than for Black applications in the pooled sample. The corresponding difference in the balanced sample is 2.2 percentage points (again 9%). These effects are driven primarily by gaps in the probability of contact by voicemail. Appendix Figure A2 shows race-specific Kaplan-Meier estimates of contact rates and hazards by days since an application was sent. Thirty days after submission, Black and white contact rates differ by 2 percentage points and contact hazards have equalized across groups. We therefore focus on 30-day contact rates for the remainder of the analysis.

Parent income, education, and other features of family background vary across distinctive names within race and gender groups (Bertrand and Mullainathan, 2004; Fryer Jr and Levitt, 2004; Gaddis, 2017). Appendix Figure A3 assesses whether employers respond to this variation by estimating separate contact rates for each first name. We fail to reject that first names have no causal effect on contact probabilities within each race-by-sex category ($p \geq 0.24$). A corresponding analysis of last names, depicted in Appendix Figure A4, also fails to reject the absence of a causal effect of names on contact rates within each race category ($p \geq 0.13$). These findings suggest that the primary effect of distinctive names is to convey race and gender to the employer.

While the overall contact rate fluctuated during the course of our study, Black applicants faced a consistent contact penalty relative to white applicants. Figure 2 shows monthly Black and white contact rates (left axis) along with the percentage gap between the rates (right axis). Contact rates fell between October 2019 and February 2020 as hiring for seasonal jobs concluded. We paused the experiment from March to August 2020 due to the COVID-19 pandemic. Contact rates were variable in the months after the experiment resumed, and sharply elevated in the final wave of our study as many states eased restrictions in the wake of widespread vaccine distribution. The measured contact rate for white applicants exceeded that for Black applicants in 12 of 13 months of the study, and we cannot reject at the 5% level that either the level or percentage contact gaps between white and Black applicants were constant across the study's five waves (or 13 months).

Our finding of a contact penalty for Black applicants corroborates a large body of evidence from resume correspondence studies reviewed in Bertrand and Duflo (2017). The 9% proportional contact gap in our study is somewhat smaller than corresponding estimates from previous work. For example, a meta-analysis by Quillian et al. (2017) concludes that white applicants typically receive 36% more callbacks than Black applicants in recent U.S. correspondence experiments. One potential explanation for the smaller proportional effect in our study is that larger firms exhibit less severe discrimination, as reported in a Canadian correspondence experiment described in Banerjee, Reitz and Oreopoulos (2018). On the other hand, the 2.1 percentage point average contact gap between white and Black applicants in our experiment aligns closely with the findings of other recent studies. For example, Nunley et al. (2015) report an average contact gap between white and Black applicants of 2.6 percentage points (17% of the Black mean), while Agan and Starr (2018) report a contact gap of 2.4 percentage points (23% of the Black mean). The lower proportional gap in our experiment is a consequence of the higher overall contact rate for our applications combined with a similar level gap in contact rates.

Our study randomized multiple protected applicant characteristics in addition to race. To summarize the overall effects of all randomized characteristics, Table 2 reports estimates of simple models of employer contact. Column (1) shows the results of fitting a linear probability model for employer contact as a function of race, sex, age, club membership, and pronouns, controlling for associate degrees, region indicators, and wave indicators. Consistent with the mean differences in Table 1, Black applications are contacted 2.1 percentage points less often than whites, a highly statistically significant difference ($p < 10^{-32}$). The corresponding estimate from a logit specification implies that Black applications face roughly 12% lower odds of a callback.

In contrast to the effect of race, the estimated average effect of sex is small and statistically insignificant. Table 2 shows that the difference in contact rates for male and female applicants is almost exactly zero, and we can reject average contact gaps of roughly 0.6 percentage points or larger in absolute value. This result is consistent with previous studies showing mixed or zero average impacts of sex on employer callbacks in the U.S. and elsewhere (Nunley et al., 2015; Baert, 2018).

We find a modest contact penalty for older applicants. The third row in Table 2 reports a statistically significant gap of 0.6 percentage points between contact rates for applicants under and over age 40. The estimate for the balanced sample is similar in magnitude but statistically insignificant. As shown in Appendix Figure A5, the probability of an employer contact declines modestly but monotonically with age, and we can reject the hypothesis that callback rates are constant across quintiles of applicant age at marginal significance levels ($p = 0.052$). Our findings for age confirm the result of Neumark, Burn and Button (2018) that age discrimination is present in the U.S. labor market, though the magnitude of age effects is somewhat smaller in our experiment.

We find limited evidence of effects of sexual orientation and gender identity, though we have less statistical precision to detect effects of these attributes than for race, gender, and age. The estimated effect of LGBTQ clubs is small and statistically insignificant in both the full and balanced samples. Gender-typical pronouns are associated with a marginally significant contact penalty of 1.3 percentage points, but this estimate is not significant in the balanced sample. Gender-neutral pronouns are associated with a comparably sized penalty that is statistically insignificant in the full sample but marginally significant in the balanced sample. Standard errors for LGBTQ clubs and pronouns are roughly three times as large as for race, a consequence of the fact that fewer than 10 percent of resumes were assigned these characteristics. We can, however, reject the 4.2 percentage point effect of LGBTQ clubs reported by Tilcsik (2011) for an earlier sample of jobs and employers. We also find no effect of listing an associate degree, a null result that is consistent with the findings of Deming et al. (2016) for non-selective jobs.

A large literature emphasizes the "intersectionality" of race and gender discrimination (Crenshaw, 1989, 1990). Table 3 investigates such interactions by comparing the effects of resume characteristics for white and Black applicants. Female names generate a marginally significant increase in contact rates for white applicants and a marginally significant decrease for Black applicants. The difference between these effects is a statistically significant 1.4 percentage points, implying that the effect of a female name is more positive for whites (or, equivalently, that the penalty for a Black name is larger for women). We also find evidence of an interaction between race and LGBTQ club status: while white applicants face a contact penalty of 1.6 percentage points for listing membership in an LGBTQ club, Black applicants receive a small, statistically insignificant, contact bonus. This difference is large enough to eliminate the contact penalty for Black names among applications listing LGBTQ club membership. While we find insignificant differences in effects for several other attributes, a joint test rejects the null hypothesis of no interaction effects across all dimensions in Table 3 ($p = 0.035$), suggesting that the gender and LGBTQ interactions are not an artifact of statistical noise.

6 Variation in Discrimination Across Firms

A central objective of our study is to measure heterogeneity across firms in the effects of protected characteristics on contact rates. If all firms have the same expected contact gap, a job seeker will have little scope to evade discrimination by redirecting their search towards less biased employers. Likewise, regulators at the EEOC or OFCCP would have little to learn from the parent company of an establishment about whether that establishment is likely engaged in discrimination.

In what follows, we use a variety of methods to document that racial and gender contact gaps vary widely across employers and are spatially and temporally stable, suggesting that the organizational structure of employment is in fact highly informative about discrimination at particular establishments. Before doing so, however, we first clarify the statistical framework used to analyze and interpret the experimental results.

6.1 Statistical framework

Denote the realized contact gap at job $j \in \{1, ..., J_f\}$ of firm f by $\hat{\Delta}_{fj}$. For most of our analysis $\hat{\Delta}_{fj}$ measures the difference between white and Black contact rates at job j, but the same construction is used to study other binary protected characteristics such as gender. Denote by Δ_f the average causal effect of race on contact rates at jobs within firm f, and let $\hat{\Delta}_f = \frac{1}{J_f}\sum_{j=1}^{J_f} \hat{\Delta}_{fj}$ be the corresponding experimental estimate given by the white/Black difference in mean contact rates at this firm. As explained in Appendix D, the population contact gap Δ_f measures the expected difference in contact rates between white and Black resumes in our experiment when sent to an average job posted by firm f. Loosely speaking, if we had repeated our experiment many times, sampling many more jobs from the same firms, each estimated firm gap $\hat{\Delta}_f$ would tend towards its population gap Δ_f.

We are interested in characterizing the distribution of Δ_f in the finite population of 108 firms in the experiment. Below, we report simple tests for whether Δ_f equals a constant Δ for all firms, as well as tests for whether $\Delta_f \geq 0$ (or ≤ 0) for all firms, implying, for example, that all firms weakly favor white applicants. Having established the direction of discrimination, a key measure of heterogeneity in discrimination will be the variance of Δ_f. This target parameter is defined as

$$\begin{aligned}
\theta &= \frac{1}{F}\sum_{f=1}^{F}\Delta_f^2 - \left(\frac{1}{F}\sum_{f=1}^{F}\Delta_f\right)^2 \\
&= \left(\frac{F-1}{F}\right)\left\{\frac{1}{F}\sum_{f=1}^{F}\Delta_f^2 - \frac{2}{F\,(F-1)}\sum_{f=2}^{F}\sum_{k=1}^{f-1}\Delta_f\Delta_k\right\},
\end{aligned}$$

where $F = 108$ is the total number of firms.

The fundamental difficulty in estimating θ is that estimation error leads the contact gap estimates $\hat{\Delta}_f$ to be more variable across firms than their population counterparts Δ_f. Formally, the "plug-in" squared contact gap estimate $\left(\hat{\Delta}_f\right)^2$ is an upward-biased estimate of Δ_f^2. The standard error s_f of $\hat{\Delta}_f$ can be used to correct this bias. In particular, a bias-corrected estimator of θ can be written

$$\begin{aligned}\hat{\theta} &= \left(\frac{F-1}{F}\right)\left\{\underbrace{\frac{1}{F-1}\sum_{f=1}^{F}\left(\hat{\Delta}_f - \frac{1}{F}\sum_{k=1}^{F}\hat{\Delta}_k\right)^2}_{\text{plug-in}} - \underbrace{\frac{1}{F}\sum_{f=1}^{F}s_f^2}_{\text{correction}}\right\} \\ &= \left(\frac{F-1}{F}\right)\left\{\frac{1}{F}\sum_{f=1}^{F}\left(\hat{\Delta}_f^2 - s_f^2\right) - \frac{2}{F\left(F-1\right)}\sum_{f=2}^{F}\sum_{k=1}^{f-1}\hat{\Delta}_f\hat{\Delta}_k\right\}.\end{aligned}$$

Variants of this estimator have been applied to estimate effect variation in several literatures (e.g., Krueger and Summers, 1998; Aaronson et al., 2007), though typically without the adjustment factor of $\frac{F-1}{F}$.

Our analysis employs the finite-sample unbiased (squared) standard error

$$s_f^2 = \frac{1}{J_f\left(J_f-1\right)}\sum_{j=1}^{J_f}\left(\hat{\Delta}_{fj} - \hat{\Delta}_f\right)^2.$$

With this choice of s_f, $\hat{\theta}$ becomes an unbiased leave out variance component estimator of the sort proposed by Kline, Saggio and Sølvsten (2020). In particular, it can be shown that

$$\hat{\Delta}_f^2 - s_f^2 = \frac{2}{J_f\left(J_f-1\right)}\sum_{j=2}^{J_f}\sum_{\ell=1}^{j-1}\hat{\Delta}_{fj}\hat{\Delta}_{f\ell},$$

which reveals that bias correcting with s_f^2 generates an estimate of Δ_f^2 based entirely on cross-products of job-level gaps. In this sense, the bias-corrected variance can be thought of as an average covariance between jobs from the same firm, which captures the common firm component of discrimination.

Generalizing this idea, we also report a "cross-wave" estimator measuring the average covariance between firm-by-wave contact gaps $\hat{\Delta}_{ft}$ and $\hat{\Delta}_{ft'}$ for all pairs ($t \neq t'$) of waves. Because the noise in each wave's estimated contact gap is independent of the noise in each other wave, this "cross-wave" covariance estimator will also yield an unbiased estimate of θ if contact gaps are stable across time. Likewise, we report a cross-state estimator that gives the average covariance between firm-by-state contact gaps $\hat{\Delta}_{fs}$ and $\hat{\Delta}_{fs'}$ for all pairs ($s \neq s'$) of U.S. states in which we sampled jobs from firm f. The ratio of the cross-wave estimator to the bias-corrected estimator provides a measure of the temporal persistence of the firm component of discrimination, while the ratio of the cross-state estimator to the bias-corrected estimator provides a measure of the geographic stability of the firm component.

6.2 Graphical evidence of firm components

We begin with a simple graphic test for the presence and stability of firm-level variance components. As above, let $\hat{\Delta}_{ft}$ be the estimated effect of race (or another binary protected characteristic) on contact gaps at firm f of jobs sampled in wave t. Likewise, define $\hat{\Delta}_{f(t)}$ as the estimated effect of the protected characteristic for all jobs at firm f *except* those sampled in wave t. Both objects estimate Δ_f, but leaving out wave t ensures that estimation errors in the latter are uncorrelated with errors in the former.

Figure 3 compares $\hat{\Delta}_{ft}$ to its leave-wave-out mean $\hat{\Delta}_{f(t)}$. To construct the figure, averages were taken of $\hat{\Delta}_{ft}$ within vingtile of $\hat{\Delta}_{f(t)}$. If there were no variation in population contact gaps across firms these conditional means would all be centered around zero. In contrast, if there exists substantial variation in population contact gaps and there were no sampling error in our experiment, then these points would lie along the dashed 45 degree line. In practice, because $\hat{\Delta}_{f(t)}$ measures Δ_f with error, the attenuated slope of this relationship gives a measure of the signal-to-noise ratio in $\hat{\Delta}_{f(t)}$.

The strong positive relationship between firm-specific contact gaps across waves implies a persistent firm component to discrimination. The points in panel (a) of Figure 3 are clustered along a line with slope 0.36, suggesting that 36% of the variation in the leave-wave-out mean of racial contact gaps is attributable to firm effects. The fact that the slope is statistically distinguishable from zero indicates that temporally persistent firm variation is detectable in our data.

Panel (b) finds a greater slope for gender contact gaps, suggesting that 46% of the variation in the leave-wave-out mean of gender contact gaps is attributable to firm effects. Given that assignment of sex was not stratified in our experiment, we expect the error variance of estimated gender contact gaps to be greater than that of racial contact gaps. Consequently, the greater reliability of the estimated gender contact gaps suggests that the population gender gaps are more variable than the population racial contact gaps across firms. Panel (c) finds an insignificant slope for age contact gaps.

Finally, panel (d) of Figure 3 plots the relationship between racial contact gaps and leave-wave-out mean gender contact gaps. The slope of the relationship is indistinguishable from zero, suggesting that firms with contact gaps favoring women are no more or less likely to exhibit discrimination against Black applicants. Overall, the strongly significant cross-wave correlations depicted here indicate that temporally persistent firm-specific contact gaps are present for both race and gender.

6.3 Formal tests of firm components

To test formally for the significance of firm-level contact gap variation, we report a classic Pearson χ^2 test of the null hypothesis that all of the population contact gaps are equal across firms. The p-values derived from this test would be exact if each firm's sample contact gap were normally distributed and centered around its population gap with variance equal to its squared standard error s_f^2.

We are also interested in whether gaps are non-negative or non-positive for all firms, which implies a common direction of discrimination. A simple but conservative test of the null hypothesis that contact gaps are weakly positive for all firms would be to compare the minimum z-score $(\hat{\Delta}_f/s_f)$ across firms to the distribution of the minimum of 108 standard normal random variables. Because this approach would be very conservative, we instead employ the high-dimensional moment inequality testing procedure of Bai, Santos and Shaikh (2021), which drops firms with strongly positive z-scores in order to improve power while still asymptotically controlling size.

The first two columns of Table 4 reports the results of these tests. Column (1) shows that the null hypothesis that racial contact gaps are equal across firms is decisively rejected by the χ^2 test. Column (2) reveals that the null hypothesis that no firms discriminate against white applicants cannot be rejected and yields a p-value of 1.00, while the null that no firms discriminate against Black applicants is decisively rejected ($p < 0.01$). The combination of these results indicates that all firms weakly favor white applicants, but some discriminate against Black applicants more than others.

Corresponding estimates for gender reveal that the overall zero effect of perceived sex masks a significant firm component to gender discrimination. As can be seen in the second row of Table 4, the χ^2 test decisively rejects that gender contact gaps are equal across firms. In conjunction with our earlier finding of no average effect of gender, this finding strongly suggests the presence of systemic discrimination against men at some firms and against women at others. Consistent with this idea, column (2) shows that we can reject both the null hypothesis of no firms discriminating against men and the null hypothesis of no firms discriminating against women at conventional levels ($p \leq 0.05$).

The third row of Table 4 demonstrates that age discrimination also varies across firms, though less strongly than for race and gender. Column (1) shows that the χ^2 test rejects the null hypothesis of constant age discrimination across firms ($p = 0.011$). As shown in column (2), we cannot reject the hypothesis that all employers weakly favor younger applicants, but the null hypothesis that no firms discriminate against older applicants is rejected at conventional levels ($p = 0.03$).

6.4 Variance components estimates

The remaining columns of Table 4 report estimates of the standard deviation of firm-level contact gaps for race, gender, and age, calculated as the square root of the unbiased variance estimate $\hat{\theta}$. The estimates for racial contact gaps reported in the first row imply substantial dispersion in discrimination across firms. As shown in column (3), the bias-corrected estimator yields a precisely-estimated standard deviation of racial contact gaps of 1.9 percentage points. The magnitude of this gap is only slightly smaller than the mean effect of 2.1 percentage points reported in Table 2. Similarly, the cross-wave and cross-state estimators yield estimated standard deviations of 1.6 and 1.8 percentage points, respectively. The similarity of the bias-corrected, cross-wave, and cross-state estimates imply that the firm component of racial discrimination is both temporally and spatially stable.

Estimates for gender in the second row of Table 4 also show large and stable firm-level discrimination components. The bias-corrected estimator reported in column (3) yields a standard deviation of gender contact gaps of 2.7 percentage points. The cross-wave and cross-state estimators produce standard deviations of 2.9 and 2.7 percentage points, again signaling temporal and spatial stability. Consistent with the weaker evidence for firm-level variation in age discrimination reported above, the cross-firm standard deviation in the effect of age over 40 equal is smaller and equal to 1.0 percentage points. The cross-wave and cross-state estimators produce positive but small estimated firm components, suggesting modest spatial and temporal persistence in age effects.

Appendix Table A2 reports corresponding evidence on firm variation in contact gaps in LGBTQ club membership, same-gender pronouns, and gender-neutral pronouns. Our study is less powered to detect firm components along these dimensions than for race, gender, and age. The estimated variance components for the effects of LGBTQ clubs and pronouns are all statistically insignificant. Appendix Table A1 shows that patterns for all protected characteristics change little in the sample of firms present in all 5 waves of the experiment.

6.5 Effects on levels vs. proportions

Some of the variation in contact gaps documented in Table 4 may stem from overall differences in firm contact rates. To assess this possibility, we fit logit, Poisson, and linear probability models (LPMs) predicting employer contact with an intercept and a Black indicator, separately by firm. We then apply the bias-corrected estimator to estimate the variances of intercept and slope parameters across firms for each model. To determine whether firms with larger contact gaps in levels also exhibit larger proportional gaps, we report bias-corrected estimates of the correlation between LPM and logit or Poisson race coefficients, netting out the portion of the correlation due to sampling error. This exercise omits the five firms with overall contact rates below 3 percent, for which estimates of odds and ratios are unlikely to be reliable.

The logit and Poisson estimates establish that our finding of a substantial firm component to racial discrimination is not driven by functional form. As shown in columns (4) and (6) of Table 5, we find large and statistically significant cross-firm variation in logit and Poisson race coefficients, with estimated standard deviations comparable to the mean effect of race in each case. Moreover, the bottom row of Table 5 reveals that the logit and Poisson coefficients are very highly correlated with the LPM contact gap, exhibiting bias-corrected correlations of 0.89 and 0.81 respectively. This strong correlation implies that conclusions regarding which firms discriminate most are likely to be very similar when discrimination is measured in levels, odds ratios, or proportions. For the remainder of our analysis we focus on levels, which have the advantage of providing a transparent measure of total contacts lost to discrimination.

7 Alternative Groupings of Jobs

Taken together, the results of the previous section establish substantial variation across firms in their average contact gaps. In this section, we investigate how the magnitude of this variation compares to other groupings of jobs.

Table 6 reports estimates of the dispersion of population contact gaps across several alternate groupings of jobs, some of which are also groupings of firms. To maximize comparability with the firm level results reported in Table 4, we adjust for imbalance in the number of jobs per firm by weighting the job level microdata in inverse proportion to the size of each job's parent firm. As described in Appendix D, this weighting ensures that variance components from groupings that nest firms, such as industry or job portal intermediary, can be given an R^2 interpretation. In cases where job groupings that do not nest firms have explanatory power, we investigate whether these groupings are significant conditional on firm fixed effects.

7.1 State

The first panel of Table 6 reports estimates of the dispersion of population contact gaps across U.S. states. In contrast to the firm-level results in Table 4, we are unable to reject the absence of a geographic component to gender or age discrimination at even the 10% level. While geographic variation in racial discrimination can be distinguished from zero at the 5% level, the estimated standard deviation of racial contact gaps across states is only 0.9 percentage points, less than half the size of the firm component reported in Table 4.

Controlling for firm effects reduces the modest state variation in contact gaps even further. Table 7 uses the methods of Kline, Saggio and Sølvsten (2020) to decompose job-level contact gaps into components attributable to state and firm fixed effects. For both race and gender, the job-weighted standard deviations of firm fixed effects are close to the estimates from Table 4, while the standard deviations of state fixed effects are negligible. The estimated variance of state gender gap fixed effects is actually negative, suggesting that this component is very small or zero. To formally test whether the state fixed effects can be distinguished from noise we employ the high dimensional heteroscedasticity-robust testing procedure of Anatolyev and Sølvsten (2020), which yields joint p-values of 0.19 and 0.48 for the state race and gender gap fixed effects, respectively. By contrast, the null hypothesis that the firm fixed effects jointly equal zero is decisively rejected for both race and gender ($p < 0.001$). Together, these results establish that the company-level variation documented in Table 4 is not explained by differences in the spatial distribution of firms' job postings.

7.2 Industry

In contrast to the results for state, the second row of Table 6 reveals substantial dispersion in discrimination across industries. Each firm in the experiment was assigned a 2-digit SIC code, grouping together industries that only contained a single firm (see Table 10 for a list). The firm-weighted standard deviation of racial contact gaps across 2-digit industries is 1.4 percentage points, while the corresponding standard deviation of gender contact gaps is 1.9 percentage points. Age contact gaps are small and statistically insignificant. Comparing the industry-level and firm-level standard deviations, we conclude that industry effects explain roughly $(0.141/0.185)^2 \times 100 = 58\%$ of the variation in racial contact gaps and $(0.186/0.267)^2 \times 100 = 49\%$ of the variation in gender contact gaps across firms.

7.3 Job titles

The finding that industry is an important predictor of multiple dimensions of discrimination leads naturally to the question of whether the sorts of jobs posted by firms are an important predictor of contact gaps. To examine this question, job titles for each job sampled in the experiment were standardized and merged to O*Net job titles using methods described in Appendix C. To maximize statistical precision, we map the 131 standardized job titles used in our O*Net merge to 41 SOC-3 codes.[9]

The third row of Table 6 reports that the standard deviation of racial contact gaps across SOC-3 codes is 1.4 percentage points and strongly statistically significant. Gender

[9] We suspect little meaningful variation is lost from this aggregation as the bias corrected variance of racial contact gaps across SOC-3 codes is numerically indistinguishable from the bias-corrected variance across standardized job titles.

contact gaps also vary significantly across SOC-3 codes, though that variability appears to be somewhat more muted than was the case with industry. Job title heterogeneity in age contact gaps is small and statistically insignificant.

To parse the separate influence of job titles and firms, Table 7 reports a decomposition of job level contact gaps into job title and firm fixed effects. Applying the bias correction of Kline, Saggio and Sølvsten (2020), the estimated standard deviation of firm effects across jobs is 0.015, while the estimated variance of SOC-3 job title effects is negative. Using the procedure of Anatolyev and Sølvsten (2020) to test that the job title effects are jointly zero yields a p-value of 0.33, suggesting that job title effects are not a major source of variation in firm contact gaps in our experiment.[10] The firm effects, by contrast, are strongly significant ($p < 0.001$).

Job titles also explain a limited share of job level variation in contact rate gaps between male and female names: the estimated standard deviation of firm effects on gender contact gaps is 0.026, while corresponding SOC-3 job title effects exhibit a standard deviation of only 0.008. The estimated covariance between firm effects and average job title effects at the firm is small and negative. As was the case with race, the null hypothesis that firm effects on gender contact gaps are jointly zero is easily rejected ($p < 0.001$) while job title effects are jointly insignificant ($p = 0.24$).

7.4 Intermediaries

The hiring websites of many large companies are hosted by third party providers of online application systems. These intermediaries often tout their ability to promote diverse and inclusive workplaces via automated screening routines (Raghavan et al., 2020). Eighty-three of the 108 firms in our experiment used an intermediary of some sort. We create 11 intermediary categories, one of which corresponds to the 25 firms hosting their own proprietary job portals and another of which groups together intermediaries employed by a single firm.

The bottom panel of Table 6 reports that the standard deviation of racial contact gaps across these intermediary codes is only 0.006. However, this component is precisely estimated and easily distinguishable from zero ($p < 0.01$). Gender gaps may also vary somewhat across intermediaries, though this component is estimated less precisely ($p = 0.05$). As with other groupings, we lack the precision necessary to detect variation in age discrimination across intermediaries. Though intermediaries seem to predict racial contact gaps, they explain only $(0.006/0.185)^2 \times 100 = 0.1\%$ of the variation across firms. This finding suggests intermediaries are not an important mediator of employer conduct towards racially distinctive names.

[10] Recall however that the experiment only sampled entry level jobs that were easy to audit with our resume technology. It may be that job titles are an important predictor of discrimination in the broader population of jobs.

8 Job, Establishment, and Firm Predictors

We next summarize relationships between discrimination and observed employer characteristics. While such relationships may not capture the causal impacts of employer attributes on discrimination, they provide a concrete description of jobs, establishments, and firms where discrimination tends to be more or less severe. Figures 4, 5, and 6 report coefficients from regressions of contact gaps on job, establishment, and firm attributes, with results for white/Black gaps in panel (a) and estimates for male/female gaps in panel (b) of each figure. Details on the measurement of all covariates appear in Appendix C.

8.1 Job characteristics

The analysis of Section 7.3 showed that contact gaps vary substantially across job titles but that this variation is insignificant conditional on firm effects. While this suggests that variation in discrimination across job titles is mostly attributable to the identity of the parent firm, it is nonetheless of interest to ask whether the task content of job titles predicts contact gaps. Figure 4 projects job-level contact gaps onto measures of the task content of the job title, constructed based on task requirements in the O*Net following Deming (2017).

The contact penalty for Black names is more pronounced among jobs requiring customer interaction (panel (a)). This result may be an artifact of customer discrimination, the empirical importance of which has been debated (Holzer and Ihlanfeldt, 1998; Leonard, Levine and Giuliano, 2010). Jobs requiring manual skills also exhibit larger racial contact gaps. Panel (b) shows that jobs requiring social or customer interaction are more likely to favor women, while jobs requiring manual skills tend to favor men. This pattern may signal discrimination on the basis of gendered stereotypes regarding characteristically female or male tasks (Goldin, 2014; Dahl, Kotsadam and Rooth, 2021). Consistent with our earlier analysis of job title effects, including firm fixed effects renders the relationships between racial discrimination and task content jointly insignificant ($p = 0.20$). For gender the task content variables are marginally significant with firm fixed effects ($p = 0.01$), suggesting the presence of some within-firm variation in gender discrimination across job types.

8.2 Establishment characteristics

Moving to establishment-level predictors, we find that racial discrimination is unrelated to county- and block-level racial mix. Panel (a) of Figure 5 shows insignificant relationships between job-level racial contact gaps and county and block racial composition, as measured in the workplace area characteristics (WAC) file derived from the Longitudinal

Employer-Household Dynamics (LEHD) database.[11] It is worth noting, however, that many jobs in our sample did not specify an exact establishment address, so block-level data are unavailable for roughly half of establishments. Our finding of no relationship between discrimination and local racial mix contrasts with the results of Agan and Starr (2020), who show that neighborhood racial composition predicts contact gaps in a sample of jobs in New York and New Jersey. This difference may be explained by our focus on large employers or the broader set of geographies included in our sample.

Racial discrimination is more severe in geographic locations with more prejudiced populations, as proxied by measures of implicit bias and racially-charged web searches. Specifically, counties with average Implicit Association Test (IAT) scores indicating more bias against Black individuals or women (measured from Harvard's Project Implicit) tend to have larger racial contact gaps (Figure 5, panel (a), top section). Similarly, contact gaps are elevated in designated media areas (DMAs) where households submit more frequent web searches for racial epithets, a measure of prejudice developed by Stephens-Davidowitz (2014). Estimates by region show that racial contact gaps are also lower in Western states.

We see little relationship between racial contact gaps and other establishment characteristics, including log establishment employment and the fraction of managers listed in the Reference USA database that are non-white or female. Moreover, the bottom of panel (a) in Figure 5 shows that including firm fixed effects renders the establishment characteristics jointly insignificant ($p = 0.34$). Similar to our analysis of job titles, this suggests that the bivariate correlations between establishment characteristics and racial contact gaps are explained by the identity of the parent firm.

Gender contact gaps are less strongly related to workplace covariates than are racial gaps. Panel (b) of Figure 5 shows insignificant relationships between gender contact gaps and local demographics, measures of prejudice, and establishment characteristics. We do see significant negative relationships between the male/female contact gap and the block-level share of female workers as well as the share of managers that are female, suggesting that the gender composition of the establishment predicts gender discrimination. These may be chance findings given the many characteristics examined, however, as the establishment characteristics are jointly insignificant with or without firm fixed effects ($p \geq 0.35$).

[11] The WAC block-level data appear to provide an accurate measure of workplace racial composition. For a small number of the firms in our sample we were able to obtain EEO-1 records documenting the racial mix of establishments with 50 or more workers. Among the 426 establishments for which we have these data, the correlation between the EEO-1 and block-level WAC measures of the fraction of Black workers is 0.79.

8.3 Firm characteristics

Firm characteristics are stronger predictors of discrimination than job or establishment characteristics. Consistent both with Becker's (1957) classic model of discrimination and the empirical findings of Pager (2016), we find that more profitable firms are less biased against Black applicants. Specifically, the top section of panel (a) in Figure 6 reveals a significant negative correlation between firm-level white/Black contact gaps and firm profits per employee. Racial discrimination is not significantly correlated with other measures of firm performance, including sales and overall firm ratings submitted by employees on the Glassdoor (GD) platform.

Racial contact gaps are smaller at companies that previously faced more regulatory scrutiny for employment practices. As shown in the middle section of Figure 6, we see less discrimination against Black applicants at firms with more Department of Labor citations for wage and hour violations and for those subject to more employment discrimination cases. Seventy-two of the 108 firms in our experiment are federal contractors.[12] Federal contractors exhibit substantially smaller contact gaps, perhaps reflecting the stronger regulatory standards to which they are held by the U.S. government.

Measures of firm diversity suggest less racial discrimination at firms with more demographic diversity among individuals with decision-making authority but no factor is individually significant. These relationships are even weaker in a multivariate regression controlling for all of the characteristics in Figure 6, indicating that some of the apparent correlation between diversity and discrimination is explained by other firm characteristics.

The strongest negative predictor of racial discrimination in our experiment is "callback centralization," measured as the number of distinct phone numbers used by the firm to contact applicants divided by the total number of jobs with at least one callback times minus one. Since this predictor is calculated using the outcome data, to avoid any mechanical relationship between job-level callback propensities and gaps we instrument centralization among one half of each firm's jobs with centralization computed in the other half, a split sample IV strategy (Angrist and Krueger, 1995). The negative coefficient estimate suggests that firms at which hiring responsibility is more centralized are less prone to bias. Overall, the firm-level variables in Figure 6 are significant predictors of racial discrimination (joint $p < 0.001$).

As with establishment characteristics, firm-level characteristics are less correlated with gender contact gaps than with racial gaps, though we do see some evidence of a relationship between firm diversity and gender discrimination. In particular, contact gaps favor women at firms with more female managers. Consistent with the results of Bertrand et al. (2019), we find an insignificant relationship between the gender mix of a company's corporate board and gender discrimination, though the point estimate suggests a weak

[12] The federal contractor status of each firm in our experiment was obtained directly from OFCCP as part of a FOIA request.

negative correlation between board female share and the male/female gap. Again, the most predictive covariate is contact centralization, which is significantly lower at firms that favor male applicants. Though most of the firm predictors of the gender contact gap are not individually significant, the joint null hypothesis that all coefficients are zero is decisively rejected ($p < 0.001$).

9 The Distribution of Discrimination

We now investigate features of the cross-firm distribution of discrimination beyond the mean and variance by adapting the non-parametric empirical Bayes deconvolution estimator of Efron (2016) to our setting. This approach extracts an estimate of the full distribution of population contact gaps Δ_f from the observed distribution of empirical gaps $\hat{\Delta}_f$ and associated standard errors s_f. The deconvolution estimator is motivated by a hierarchical model for the firm-specific z-scores $z_f = \hat{\Delta}_f / s_f$ and their population analogues $\mu_f = \Delta_f / s_f$:

$$z_f \mid \mu_f \sim \mathcal{N}(\mu_f, 1), \quad \mu_f \sim G_\mu, \quad \text{for } f = 1, \ldots, 108.$$

The normality assumption for z_f can be justified by an asymptotic approximation with a growing number of jobs sampled for each firm. The distribution G_μ of studentized contact gaps is assumed to belong to an exponential family flexibly parameterized by a fifth-order spline. This procedure produces penalized maximum likelihood estimates of the spline parameters, yielding an implied distribution $\hat{G}_\mu$ of studentized contact gaps with corresponding density $\hat{g}_\mu$.

Assuming that s_f is independent of μ_f, we can then recover the distribution G_Δ of unstudentized contact gaps Δ_f. An estimate of the contact gap density $g_\Delta(x) = dG_\Delta(x)$ is obtained at each point x by numerically evaluating the integral

$$\hat{g}_\Delta(x) = \int \exp(-t) \hat{g}_\mu(\exp(-t)x) \hat{g}_{\ln s}(t) dt,$$

where $\hat{g}_{\ln s}$ denotes a kernel estimate of the density of log standard errors $\ln s_f$. Appendix Table A3 assesses the independence assumption by reporting coefficients from regressions of z_f on s_f, as well as regressions of the resulting squared residuals on s_f. To account for possible correlated estimation error in s_f and z_f we also report split-sample versions of these regressions that randomly partition the data for each firm and compute the z-scores and standard errors in separate half-samples. These estimates show weak relationships between standard errors and z-scores for both race and gender, suggesting that independence is a reasonable approximation.

Panel (a) of Figure 7 displays the deconvolved density of contact gaps between white and Black applicants, while panel (b) reports the density of gaps between male and female applicants. The penalization parameter of the first-step maximum likelihood procedure is calibrated to yield a variance matching the bias-corrected estimate in Table 4.[13] In panel (a) we restrict the support of the density of racial contact gaps to rule out discrimination against whites—a shape constraint we showed earlier cannot be rejected by our data.[14] For comparison with the estimated densities, the background of Figure 7 also reports histograms of firm contact gap estimates $\hat{\Delta}_f$. As a result of the noise in these estimates, the contact gap distributions implied by the histograms are substantially more dispersed than the deconvolved distributions.

The deconvolved density of racial contact gaps reveals a skewed distribution with a thick tail of extreme discriminators that favor white applicants by more than 5 percentage points. This density can be approximated closely by a log-normal distribution with the same mean and variance. Panel (b) shows that the estimated distribution of population gender gaps is nearly symmetric around zero and heavily leptokurtic. This distribution turns out to be even more strongly peaked about its mode than a Laplace distribution with identical mean and variance, indicating that many companies exhibit very little gender bias, while a small number of severe discriminators are biased in each direction.

The distributional estimates for both race and gender imply that a large share of discrimination is driven by a small group of highly discriminatory firms. Figure 8 summarizes the concentration of discrimination by plotting the Lorenz curve implied by the deconvolved density $\hat{g}_\Delta$. The Lorenz curve for race measures the share of the total contact gap between white and Black applications in the experiment attributable to firms below each percentile of Δ_f. Since gender discrimination operates in both directions, the gender curve summarizes concentration of the absolute contact gap $|\Delta_f|$.

The discrimination Lorenz curves are strongly bowed away from the 45 degree line, implying that discrimination is highly concentrated in particular firms. For example, the race Lorenz curve shows that firms in the top quintile of discrimination are responsible for 46% of lost contacts to Black applicants in our study, while firms in the bottom quintile are responsible for less than 5% of lost contacts. The gender curve exhibits even more inequality, with firms in the top quintile responsible for 57% of aggregate absolute gender differences in the experiment.

The area between each Lorenz curve and the 45 degree line gives the Gini coefficient, which ranges from 0 (perfect equality) to 1 (perfect concentration). For race, the Gini coefficient is roughly 0.40, which is nearly as large as estimates of the Gini for modern

[13] As Efron and Tibshirani (1996) note in a closely related context, imposing such moment constraints can provide an attractive balance between local adaptivity and respecting certain global properties of the density.

[14] For race, we set the support of G_μ to $[0, \max_f(z_f) + 0.5]$. The support of G_Δ is assumed to be $[0, \max_f(z_f) \max_f(s_f)]$. For gender, we assume the supports of G_μ and G_Δ are $[\min_f(z_f) - 0.5, \max_f(z_f) + 0.5]$ and $[\min_f(z_f) \max_f(s_f), \max_f(z_f) \max_f(s_f)]$, respectively. A deconvolved density of racial contact gaps that does not impose the positive support restriction is reported in Figure A9.

U.S. income inequality. For gender, the Gini coefficient is 0.54, substantially higher than Gini income estimates in the U.S. and roughly comparable to Brazil's level of income inequality.[15]

10 Firm-Specific Estimates

We have established that firms differ substantially in their average contact rates by race and gender. We now turn to studying whether it is possible to determine which particular firms are engaged in discrimination. Our analysis of firm-specific discrimination leverages empirical Bayes (EB) methods that "borrow strength" from the full set of firms in the experiment to improve estimates of contact gaps at specific firms.

10.1 Posterior estimates

The EB framework treats the mixing distributions estimated in Section 9 as priors to construct posterior distributions for each individual firm. The EB posterior mean for the contact gap at firm f is given by

$$\bar{\Delta}_f = s_f \times \frac{\int x\varphi(z_f - x)\hat{g}_\mu(x)dx}{\int \varphi(z_f - x)\hat{g}_\mu(x)dx},$$

where φ denotes the standard normal density. The posterior mean $\bar{\Delta}_f$ constitutes a best (i.e., minimum mean squared error) predictor of the population contact gap Δ_f when treating the estimated population distribution $\hat{G}_\Delta$ as background knowledge. For comparison, we also compute linear shrinkage estimates obtained by taking a precision-weighted average of the estimated gap and grand mean:

$$\tilde{\Delta}_f = w_f\hat{\Delta}_f + (1 - w_f)\frac{1}{108}\sum_{f'=1}^{108}\hat{\Delta}_{f'}.$$

The weights are given by $w_f = \frac{\hat{\theta}}{s_f^2+\hat{\theta}}$, where $\hat{\theta}$ is the square of the between-firm standard deviation estimate reported in Table 4. Estimators of this sort are used heavily in economics (e.g., Kane and Staiger, 2008; Chetty, Friedman and Rockoff, 2014; Angrist et al., 2017; Chetty and Hendren, 2018; Abaluck et al., 2021) but correspond to EB posterior means only when G_Δ is assumed normal. Even if the prior distribution is not normal, however, $\tilde{\Delta}_f$ retains an interpretation as a best linear predictor of the population gap Δ_f given the estimated gap $\hat{\Delta}_f$.

The EB posterior means are highly variable across companies, implying that the experiment contains substantial information about the behavior of individual firms. Figure

[15] See `https://data.worldbank.org/indicator/SI.POV.GINI/`.

A8 compares the distributions of observed contact gaps $\hat{\Delta}_f$, EB posterior means $\bar{\Delta}_f$ and linear predictions $\tilde{\Delta}_f$, and the estimated prior distribution $\hat{G}_\Delta$. The distribution of posteriors is more compressed than the observed contact gaps $\hat{\Delta}_f$ or the deconvolved prior distribution $\hat{G}_\Delta$, reflecting shrinkage due to the noise in the observed gaps. Unlike the observed contact gaps, the posterior means are strictly positive, inheriting the non-negativity constraint placed on the prior distribution. In contrast, roughly 12% of the linear shrinkage estimates are negative, a consequence of the symmetric implicit normal prior. The upper tail of the distribution of linear shrinkage estimates is more compressed than is the distribution of empirical Bayes posterior mean estimates, which reflects that the roughly log-normal shape of our estimated prior $\hat{G}_\Delta$ exhibits a fat tail of heavy discriminators. The EB posterior accounts for this fat tail by applying less shrinkage to extreme positive contact gaps. Overall, 46 firms have posterior mean racial contact gaps greater than the average gap of 2 percentage points in the experiment.

The posterior mean racial contact gaps vary systematically across industries. Figure 9 reports mean values of $\bar{\Delta}_f$ and $\tilde{\Delta}_f$ by 2-digit industry. Racial discrimination is estimated to be particularly severe among firms in customer-facing sectors. The posterior mean contact gap averages 4.2 percentage points among the eight firms in the auto dealers and services sector (SIC 55), 2.9 percentage points for the five firms in the eating and drinking sector (SIC 58), and 2.7 percentage points for the four apparel firms (SIC 56) in the experiment. By contrast, the posterior mean racial contact gap averages only 0.9 percentage points among the two engineering services firms (SIC 87), 1 percentage point among the five banking and credit firms (SICs 60-61), and 1.1 percentage points among securities brokerages (SIC 62) and freight and transport firms (SICs 42-47) in our experiment.

Posterior estimates of gender discrimination also vary across industries. Discrimination against men appears concentrated in the apparel sector, where distinctively male names face a severe contact disadvantage of 6.5 percentage points. Discrimination against women appears most pronounced among the two firms in the wholesale durable sector (SIC 50) where distinctively female names face an average contact disadvantage of 3.9 percentage points. In line with the strong peak in the prior distribution around zero reported in panel (b) of Figure 7, however, many sectors are estimated to exhibit trivially small gender contact gaps. Indeed, the three firms in the business services sector (SIC 73) exhibit an average posterior mean gender contact gap of zero.

Figure 10 plots coefficients from the projection of industry characteristics (normalized to have standard deviation 1) onto the firm posterior mean contact gaps. Firms estimated to favor white applicants reside in industries with somewhat lower Black employment shares and female employees concentrated in non-management positions, but the relationships are only marginally significant. By contrast, firms estimated to favor male applicants lie in sectors with sharply lower female employment shares, higher unexplained gender wage gaps, and Black employees concentrated in non-management positions. One potential interpretation of these patterns is that job seekers know that certain sectors (e.g. women's apparel) discriminate on the basis of gender, which allows them to sort away from biased jobs, mitigating to some extent the burden of discrimination as in Becker's (1957) classic model. Racial discrimination, by contrast, may be more difficult to detect based on industry, leading to less pronounced sorting patterns and a larger burden on job seekers when search is costly (Black, 1995; Bowlus and Eckstein, 2002).

10.2 Guarding against false discoveries

While the posterior mean estimates of the previous section provide a best guess of the contact gap at each firm, it is possible that some firms with large posterior mean contact gaps have true population gaps of exactly zero. The question of whether a firm's contact gap is exactly zero has direct legal relevance as the Civil Rights Act prohibits *any* discrimination based upon protected characteristics. To assess the conclusions that can be drawn about which employers are discriminating at all, we next consider a related class of empirical Bayes methods that aims to limit false discoveries.

For each firm in our experiment, we can assign a p-value $\hat{p}_f$ to the null hypothesis that the firm's population contact gap is zero by comparing the firm's z-score to the appropriate tail of a t-distribution.[16] Histograms of the resulting p-values for the null that firm specific contact gaps equal zero appear in Figure 11. Panel (a) of the figure reports one-tailed tests of the null of no discrimination against Black applicants, while panel (b) reports two-tailed tests of the null that racial contact gaps are exactly zero. Panel (c) reports two-tailed tests that gender contact gaps are zero.

If all firms had racial and gender contact gaps equal to zero, we would expect all three histograms to be uniformly distributed. In practice, we see substantial bunching of the $\hat{p}_f$ at small values. For example, 31 firms (28.7%) have one-tailed p-values for the null of no racial discrimination below 0.05, while 14 firms (13.0%) have two-tailed $\hat{p}_f$ below 0.05 for the null of no gender discrimination. Applying Tukey's "higher criticism" criterion (Donoho and Jin, 2004), even the modestly elevated share of small p-values for gender discrimination indicates a significant departure from uniformity at the 5% level, as $\sqrt{108} \times \left(\frac{0.13-0.05}{\sqrt{0.05\times 0.95}}\right) \approx 3.81 > 1.96$. Clearly some firms are discriminating, but which ones?

Recall that in the deconvolution analysis of the previous section, we assumed the population contact gaps were drawn from a continuous distribution G_Δ. Suppose instead that a proportion $\pi_0 \in [0, 1]$ of all firms have population contact gaps exactly equal to

[16] We set the degrees of freedom of the t-distribution equal to the number of jobs at the firm minus one.

zero. Let $F_{\hat{p}}$ denote the distribution of empirical p-values. By Bayes' rule, the posterior probability that a firm with a $\hat{p}_f$ less than $p \in (0,1]$ has a contact gap of exactly zero can be written

$$FDR\left(p\right) = \frac{\Pr\left(\hat{p}_f < p \mid \Delta_f = 0\right)\pi_0}{\Pr\left(\hat{p}_f < p\right)} = \frac{p\pi_0}{F_{\hat{p}}\left(p\right)},$$

where the second equality follows from the p-values being uniformly distributed among the sub-population of firms with zero contact gaps. $FDR\left(p\right)$ has a frequentist interpretation as the expected proportion of null hypotheses with p-values less than p that are true, a quantity known in the multiple testing literature as the False Discovery Rate (Benjamini and Hochberg, 1995).[17]

Storey (2002) introduced the idea of deciding on null hypotheses according to their "q-values," which can be thought of as empirical Bayes analogues of p-values. The q-value for rejecting all nulls with p-values less than $\hat{p}_f$ is

$$\hat{q}_f = \widehat{FDR}\left(\hat{p}_f\right) = \frac{\hat{p}_f\hat{\pi}_0}{\hat{F}_{\hat{p}}(\hat{p}_f)},$$

where $\widehat{FDR}\left(p\right)$ is an estimator of false discovery rates based on the empirical distribution of p-values.[18] If $\widehat{FDR}\left(p\right)$ were a consistent estimator of $FDR\left(p\right)$ then classifying all firms with q-values less than 0.1 as discriminators should yield a False Discovery Rate of 10% – i.e., we should expect 10% of these firms to actually have zero contact gaps.

The primary difficulty in computing a suitable estimator of $FDR\left(p\right)$ is that the proportion π_0 of nulls that are true is not point identified. The testing procedure of Benjamini and Hochberg (1995) effectively sets $\pi_0 = 1$. Efron et al. (2001) note that a more informative upper bound on π_0 is given by the minimal density $\min_{p\in(0,1]} f_{\hat{p}}(p)$ of the p-values. The minimum should be achieved near the point $p = 1$, as large p-values are more likely to be generated by nulls that are true. Building on this idea, Storey (2002) proposed the tail density estimator

$$\hat{\pi}_0\left(\lambda\right) = \frac{\sum_{f=1}^{108} 1\left\{\hat{p}_f > \lambda\right\}}{\left(1-\lambda\right)108},$$

where $\lambda \in [0,1)$ is a tuning parameter governing how deep to look in the right tail of empirical p-values. Any value of λ provides an upper bound on π_0. Larger values of λ will tend to yield less conservative bounds but more sampling variability. We use the automated bootstrap procedure of Storey et al. (2015) to balance variance against conservatism in our choice of λ.[19] To assess the degree of uncertainty in our estimate, we

[17] See Storey (2003) and Efron (2016) for more on empirical Bayes interpretations of false discovery rates. We have implicitly assumed that at least one firm has a $\hat{p}_f$ less than p.

[18] In practice we follow Storey (2002, 2003) in estimating $\hat{q}_f$ as $\min_{t\geq\hat{p}_f} \widehat{FDR}(t)$. This ensures that q-values are non-decreasing for nested rejection thresholds.

[19] For any choice of λ, however, the probability limit of $\hat{\pi}_0\left(\lambda\right)$ should lie weakly above the true π_0.

also report the upper limit of a non-parametric confidence interval for π_0 developed by Armstrong (2015).

10.3 Which firms discriminate?

Figure 11 reports choices of λ and the estimated tail density $\hat{\pi}_0(\lambda)$ for both one and two-tailed tests of racial discrimination. As expected, the $\hat{\pi}_0(\lambda)$ correspond roughly to the right asymptote of the plotted discrete density estimates. Super-imposed on Figure 11 are estimates of the Local False Discovery Rates (LFDRs; Efron et al., 2001) implied by setting $\pi_0 = \hat{\pi}_0(\lambda)$. LFDRs give posterior estimates of the probability that a null hypothesis is true given its p-value. The mean LFDR below a threshold p-value $\hat{p}_f$ gives an approximation to $\hat{q}_f$.[20]

For one tailed tests we estimate that $\pi_0 \leq 0.39$; i.e., that at least 61% of firms discriminate against Black applications. Unsurprisingly, allowing for bi-directional racial discrimination dissipates power, leading to an upper bound on π_0 of 0.54. Table 8 provides a sensitivity analysis involving a few other estimates of π_0. Computing the p-values via randomization inference tends to yield more very small p-values, resulting in a correspondingly smaller estimate of π_0.[21] Estimating π_0 with a cubic spline, as in Storey and Tibshirani (2003), yields slightly larger estimates of π_0. The final panel of the table reports the upper limit of a 95% confidence interval on π_0. For one-sided tests as few as 42% of firms may be discriminating against Black applicants, while under two-tailed tests the share discriminating may be as low as 30%.

The resulting q-values imply that many individual firms can be reliably detected as discriminating against Black applicants. In our benchmark specification 23 firms have q-values less than 0.05 (Table 8, top panel, first column). Table 9 lists industry, federal contractor status, contact gap estimates, posterior means and quantiles, and p- and q-values for this set of companies (with firm names suppressed). Since a decision rule based on $q = 0.05$ limits the False Discovery Rate to at most 5%, we should expect at most $23 \times 0.05 = 1.15$ errors if these 23 firms were classified as discriminating against Black applicants. Interestingly, the firm with the largest q-value has a posterior mean contact gap of 1.8pp and a posterior 5th percentile gap of 0.75pp, indicating that if the deconvolved distribution $\hat{G}$ is taken as a prior, one can be confident that a non-trivial amount of discrimination is taking place at this firm.

Conclusions regarding the detection of individual employers that discriminate against Black applicants are remarkably robust to the method used to bound π_0. In fact, if we

[20] Letting $f_{\hat{p}}$ denote the density of observed p-values, we can define $LFDR\,(p) = \pi_0/f_{\hat{p}}(p)$. It is straightforward to verify that $FDR\,(p) = \int_0^p f_{\hat{p}}\,(b)\,LFDR\,(b)\,db/F_{\hat{p}}\,(p)$. Because we use a kernel smoother to estimate $f_{\hat{p}}$, the running average of LFDR estimates does not numerically match $\hat{q}_f$ in sample.

[21] Randomization based tests avoid reliance on asymptotics but evaluate the "sharp" null that none of the firm's contact decisions were influenced by protected characteristics. See Ding (2017) for further discussion of how to interpret such tests.

use estimates of p-values and $\hat{\pi}_0$ based on randomization inference, the 23 firms assigned racial discrimination q-values less than 0.05 in our baseline analysis have an average LFDR of only 0.025, suggesting the False Discovery Rate for this collection of firms may actually be 2.5% or less. When π_0 is set to the upper limit of its 95% confidence interval – an extremely conservative choice – 20 firms have q-values below 0.05 (Table 8, bottom panel, first column). This prior insensitivity arises because many firms have very small p-values, as shown in Table 9.

Consistent with the posterior mean estimates in Figure 9, we find a clear industry pattern among firms with low q-values for discrimination against Black applicants. As shown in Table 10, firms detected as discriminating against Black names are highly concentrated in the auto dealers and services sector, where 6 of the 8 firms in our experiment have q-values below 0.05. The mean LFDR in this sector is 0.15, implying that at least 85% of the firms in this industry discriminate against Black applicants. Other sectors with a high concentration of racial discrimination include other retail (SIC 59), where 3 of the 7 firms have q-values below 0.05, and furnishing stores (SIC 57), where 2 of 4 firms have low q-values. Mean LFDRs are substantially higher than 0.05 in these sectors, indicating that the firm-specific p-values remain somewhat dispersed within industry. Notably, 8 of the 23 firms with q-values less than 0.05 are federal contractors, including the firm with the highest posterior mean level of racial discrimination.

To further compare results based on posterior means and q-values, Appendix Figure A10 plots the posterior mean racial contact gaps ($\bar{\Delta}_f$) from the previous section against the $\hat{q}_f$ from our preferred specification. Bracketing the posterior means are 95% empirical Bayes credible intervals (EBCIs) connecting each firm's posterior 2.5th percentile contact gap to its posterior 97.5th percentile. If the prior $\hat{G}_\Delta$ were estimated without error then 95% of the population contact gaps would be expected to lie within these confidence intervals. The lower limit of each EBCI is positive because the estimated prior imposed that racial contact gaps are almost surely positive. By contrast, the q-values were derived under the assumption that 39% of firms have contact gaps of exactly zero. As expected the posterior mean contact gaps are generally decreasing in $\hat{q}_f$ but the relationship between the two measures is not perfectly monotone.

As a result of the higher concentration of gender contact gaps near zero, it is more difficult to detect individual firms discriminating on the basis of gender than on the basis of race. Panel (c) of Figure 11 shows the distribution of p-values derived from tests that gender contact gaps are zero. Here the Storey et al. (2015) procedure produces an upper bound on π_0 of 0.83, implying that at least 17% of firms discriminate on the basis of gender. Moreover, the 95% confidence interval on π_0 extends to 1, suggesting that we cannot reject the null that none of the firms discriminate based upon gender. This conclusion is clearly at odds with our earlier higher criticism calculation, not to mention the tests presented in Table 4, which decisively rejected the null that gender contact gaps are equal across firms. This discrepancy likely arises because the Armstrong (2015) test is designed to have good power properties in settings where π_0 is small, which seems not to be the case here.[22] Likewise, the 95% confidence interval for the proportion of firms not discriminating against older applicants also includes 1, which is unsurprising given that the tests reported in Table 4 detected only modest firm heterogeneity in age discrimination.

These high estimated bounds on π_0 lead to high lower bounds on the posterior probabilities of gender discrimination for most firms. Consequently, Table 8 shows that only one firm has a q-value for gender discrimination below 0.05.[23] Table 10 indicates that this company is in the apparel sector. Based on its posterior mean, this apparel store is discriminating against men. Interestingly, the same store also has a q-value below 0.05 for racial discrimination. While the apparel sector (SIC 56) has a large average posterior mean contact gap favoring women, the mean LFDR in the sector is relatively high, suggesting industry membership is not, in itself, dispositive of gender discrimination.

10.4 Prevalence vs. severity

Having established with high posterior certainty that 23 firms favor white applicants on average, we now examine whether these firms' racial contact gaps could have been generated by a small minority of discriminating jobs. This distinction between the prevalence and severity of racial discrimination is arguably pertinent to the legal notion of systemic discrimination as a widespread pattern of organizational behavior. Kline and Walters (2021) show that the share of jobs that discriminate is not point identified in audit designs sending a small number of applications to each job. Consequently, we rely on a simple bounding approach to assess the prevalence of discrimination across jobs within firms.

To formalize the notion of job-level discrimination prevalence, it is convenient to again work with a mixture representation. Suppose that a proportion $1 - \phi_f$ of the jobs at firm f have contact gaps of exactly zero.[24] With this notation, the firm-wide mean contact gap can be written $\Delta_f = \phi_f \dot{\Delta}_f$, where $\dot{\Delta}_f$ gives the average contact gap among discriminating jobs within firm f. Here $\dot{\Delta}_f$ provides a measure of discrimination severity, while ϕ_f indexes the prevalence of discrimination.

[22] As Armstrong (2015) notes, his procedure "looks at the larger ordered p-values in order to achieve adaptivity to the smoothness of the distribution of p-values under the alternative in a setting where π may not be close to 1."

[23] Note that this firm has a q-value below 0.05 even when $\bar{\pi}_0 = 1$. This occurs because $\hat{p}_f$ is well below $\hat{F}_{\hat{p}}(\hat{p}_f)$, so that $\hat{q}_f$ is small even when plugging in an upper bound on π_0 of unity.

[24] One reason that a particular job may not discriminate is that its population contact rate may be zero, for instance, because the job may have already been filled. Consequently, even a firm with a practice of always discriminating in hiring might, by this definition, exhibit a $\phi_f < 1$.

The variance of job level contact gaps at firm f can be written

$$\sigma_f^2 = \phi_f \dot{\sigma}_f^2 + \phi_f(1-\phi_f)\dot{\Delta}_f^2,$$

where $\dot{\sigma}_f^2$ denotes the variance of contact gaps among discriminating jobs. Note that $\sigma_f^2 \geq \phi_f(1-\phi_f)\dot{\Delta}_f^2$, which binds with equality when all discriminating jobs exhibit equal population contact gaps. Substituting this bound into the expression for Δ_f and rearranging yields the following lower bound on discrimination prevalence at firm f:

$$\phi_f \geq \Delta_f^2/(\sigma_f^2 + \Delta_f^2).$$

A simple rule of thumb emerges from this expression: if the mean level of discrimination is roughly equal to its standard deviation – as was found for the distribution of racial contact gaps across firms – then prevalence must be at least one half. Interestingly, the density based prevalence bounds reported in Table 8 were only slightly above one half, suggesting this moment-based bound sacrifices little identifying information when applied to firm-wide average gaps.

An unbiased estimate of the variance of job-level gaps can be computed by taking the covariance between contact gaps for the first and last two application pairs sent to each job. Applying this approach, Appendix Table A4 reports that the standard deviation of contact gaps across all jobs in the experiment is 0.071. The mean gap across jobs is 0.020 with associated standard error of 0.002. Consequently, the lower bound prevalence is estimated to be $\frac{(0.020)^2-(0.002)^2}{(0.020)^2-(0.002)^2+(0.071)^2} \approx 0.07$, indicating that at least 7% of jobs in the experiment as a whole discriminate against Black names.

We can conduct a corresponding calculation at each firm, using $\hat{\Delta}_f^2 - s_f^2$ as a bias corrected estimate of each firm's Δ_f^2. Figure 13 illustrates these firm specific estimates, which are quite noisy, ordered by the firm's q-value. As expected, firms with lower q-values tend to have higher job-level prevalence bounds. To reduce sampling error, the solid line plots the average bound among jobs at firms with q-values below a threshold level. Firms with $q < 0.1$ for example have a lower bound prevalence of 18%. The 23 firms with $q < .05$ exhibit a prevalence of at least 20%, suggesting that discrimination against Black names is widespread among the establishments that comprise these firms.

11 Detection Possibilities

We now bring together the two empirical Bayes classification schemes considered earlier to investigate the "price" paid for classifying firms according to their q-values rather than their posterior means. To frame this trade-off, it is useful to tie our discussion to the decision problem faced by a hypothetical auditor charged with deciding whether to investigate the firms in our study. While this analysis is stylized, it is worth noting that the EEOC, OFCCP, and several local organizations, such as the New York City Commission on Human Rights, proactively investigate systemic discrimination on an ongoing basis. Statistical evidence is a legally recognized basis for such decisions.[25]

11.1 The auditor's problem

Consider an auditor concerned with racial discrimination who can launch investigations into the conduct of any of the firms in our experiment at cost $c \in (0,1)$. Let $\delta_f \in \{0,1\}$ be an indicator for the decision to launch an investigation into firm f and $\mathcal{D}$ the collection of these indicators.

We consider two potential specifications of the auditor's preferences that differ in whether she is concerned with the intensive or extensive margin of discrimination. These two objectives can be written as functions of the unobserved racial contact gaps $\{\Delta_f\}_{f=1}^{108}$, each of which is assumed to lie in the unit interval:

$$\begin{aligned} U^i(\mathcal{D}) &= \sum_{f=1}^{108} \delta_f \left(\Delta_f - c\right), \\ U^e(\mathcal{D}) &= \sum_{f=1}^{108} \delta_f \left(1\{\Delta_f > 0\} - c\right). \end{aligned}$$

An auditor with preferences given by U^i would like to investigate every firm with $\Delta_f > c$, while an auditor with preferences given by U^e seeks to investigate every firm with $\Delta_f > 0$. The latter objective arguably reflects U.S. employment law, which prohibits any discrimination on the basis of race. One can also think of U^e as capturing an extreme form of risk aversion regarding the unobserved racial contact gaps.[26]

The auditor must rely on the experimental evidence $\mathcal{E} = \left\{\hat{\Delta}_f, s_f\right\}_{f=1}^{108}$ to make decisions regarding which firms (if any) to investigate. Given a prior G over the distribution of population contact gaps, the auditor's expected utility under these two preference

[25] For example, the U.S. Department of Labor's Administrative Review Board ruled in Office of Federal Contract Compliance Programs, U.S. Department of Labor v. Bank of America (2016) that "the more severe the statistical disparity, the less additional evidence is needed to prove that the reason was race discrimination. Very extreme cases of statistical disparity may permit the trier of fact to conclude intentional race discrimination occurred without needing additional evidence." See Office of Federal Contract Compliance Programs, U.S. Department of Labor v. Enterprise RAC Company of Baltimore, LLC (2019) for a similar ruling by the Office of Administrative Law Judges.

[26] Both utility functions can be viewed special cases of the more general preference scheme $U(\mathcal{D}) = \sum_{f=1}^{108} \delta_f \left(\Delta_f^{1/p} - c\right)$, where $p \geq 1$ governs the auditor's risk aversion. When $p = 1$, the auditor is risk neutral and $U = U^i$. As $p \to \infty$, U approaches U^e.

schemes can be written

$$\begin{aligned} \mathbb{E}_G[U^i(\mathcal{D})|\mathcal{E}] &= \sum_{f=1}^{108} \delta_f \left(\bar{\Delta}_f(G) - c\right), \\ \mathbb{E}_G[U^e(\mathcal{D})|\mathcal{E}] &= \sum_{f=1}^{108} \delta_f \left(1 - LFDR_f(\pi_0) - c\right), \end{aligned}$$

where $\Delta_f(G) = \mathbb{E}_G[\Delta_f|\mathcal{E}]$ is the posterior mean contact gap for firm f, $LFDR_f(\pi_0) = \Pr_G(\Delta_f = 0|\mathcal{E})$ is the posterior probability that firm f is not discriminating, and $\pi_0 = G(0)$ is the prior probability of non-discrimination.

If, based on $\mathcal{E}$, an auditor with preferences U^i were to settle on beliefs over contact gaps coinciding with the deconvolved distribution $\hat{G}_\Delta$, then she would investigate all firms with empirical Bayes posterior means $\bar{\Delta}_f$ exceeding c. If the auditor instead believes population contact gaps are normally distributed with a variance equal to that reported in Table 4, she will investigate all firms with linear shrinkage estimates $\tilde{\Delta}_f$ exceeding c.

The decision problem is somewhat trickier for an auditor with preferences U^e who is willing to entertain the possibility that a large share of firms are not discriminating at all. Recall that the probability of non-discrimination π_0 is, in general, only bounded by our experiment (Efron et al., 2001; Kline and Walters, 2021). Faced with this ambiguity, an auditor with preferences U^e might reasonably consider the largest value of π_0 consistent with the experimental evidence. Optimizing against this least favorable value $\pi_0^\dagger$ of π_0 leads the auditor to investigate all firms with $LFDR_f(\pi_0^\dagger) < 1-c$. This *minimax* decision rule coincides with a q-value based threshold, as q-values are running averages of (sorted) LFDRs.

A natural question raised by these derivations is how often a minimax auditor concerned with extensive margin discrimination would dispute the decisions of an (empirical) Bayes auditor concerned with the intensive margin of discrimination. In principle, LFDR based rankings of firm behavior can differ substantially from rankings based on posterior means (Gu and Koenker, 2020). Reassuringly, we demonstrate below that little would be lost from investigating firms based upon q-value thresholds even from the perspective of an auditor with preferences given by U^i and smooth priors given by $\hat{G}_\Delta$.

11.2 Detection possibility frontiers

Figure 12 illustrates graphically the tradeoff the auditor faces between the costs of investigating more firms and the benefits of finding additional large contact gaps. Suppose that 1,000 Black applications are sent at random to jobs equally distributed across the firms in our experiment, and contact gaps among these firms follow the estimated distribution $\hat{G}_\Delta$. The figure reports the contacts expected to be lost due to racial discrimination among investigated firms under various investigation threshold rules. The dotted 45 degree line gives the results of investigating firms at random: because the mean contact gap is 2 percentage points, investigating all the firms would "save" roughly 20 contacts per 1,000 applications, while investigating half of the firms at random would save 10 contacts.

The solid line illustrates the detection possibilities frontier available to the auditor if she observed the Δ_f without error. This infeasible frontier is simply a rescaled Lorenz curve for the distribution $\hat{G}_\Delta$. Reflecting that distribution's fat tail, the worst 20% of discriminating firms are responsible for roughly half of the lost contacts. The preferences of an auditor with objective U^i can be visualized as indifference lines with slope -1000c. An optimum occurs at a point of tangency between the indifference line and the detection frontier.

The dashed dotted line illustrates the frontier that arises when the auditor selects firms based on their posterior means $\bar{\Delta}_f$. The vertical distance between the posterior mean frontier and the true contact gap frontier reflects the cost of ranking firms according to their posterior means rather than their true contact gaps. Because the distribution of posteriors is more compressed than $\hat{G}_\Delta$, the auditor must investigate roughly a quarter of the firms based on their posterior means to isolate those responsible for half of lost contacts.

Selecting firms using the linear shrinkage estimator $\tilde{\Delta}_f$ instead of $\bar{\Delta}_f$ is estimated to entail only a small degradation of the possibilities frontier. This robustness reflects the high degree of rank correlation between the posterior mean and the linear shrinkage estimator ($\rho = 0.9$). Though the firm rankings are highly correlated across shrinkage methods, an auditor would likely choose to investigate fewer firms based on the linear shrinkage estimator, which predicts that fewer firms are engaged in severe discrimination against Black applicants.

Finally, the dashed line illustrates the frontier that arises when selecting firms based on q-values under the maintained assumption that contact gaps are distributed according to $\hat{G}_\Delta$. The expected cost of ranking firms based on their q-values, as would be optimal under preference scheme U^e, rather than their posterior means is surprisingly small, though performance degrades somewhat when more than half of the firms are investigated. Notably, the roughly 21% (23/108) of firms with q-values less than or equal to 0.05 are responsible for roughly 40% of lost contacts. Investigating the same share of firms based on posterior mean rankings would therefore be expected to yield less than 10% of additional lost contacts. Evidently, the price to be paid for control over false discoveries in our setting is fairly small. More generally, these results imply that it is possible to detect individual firms responsible for a substantial share of racial discrimination even while maintaining a tight limit on false-positive investigations of non-discriminators.

12 Conclusion

Our analysis establishes that many large U.S. employers exhibit nationwide patterns of racial discrimination that are temporally and spatially stable. Racial and gender contact gaps are highly concentrated in particular firms. We estimate that the 20% of firms discriminating most heavily against Black names are responsible for roughly half of the contacts lost to racial discrimination in our experiment. Racial discrimination appears to be widespread among the jobs posted by these firms.

In principle, the concentration of discriminatory behavior in a sub-population of employers could dampen the economy-wide consequences of discrimination, as workers can sort away from biased firms (Becker, 1957). Such a conclusion hinges crucially, however, on whether workers are aware of firm differences in average behavior. The relatively weak correlations between racial contact gaps and local demographics uncovered in our analysis give us reason to question this assumption. Rather, our impression is that the identities of the 23 firms conclusively determined to be discriminating against Black names would come as a surprise both to the companies involved and to the public at large. The identities of the companies likely discriminating on the basis of perceived sex are somewhat less surprising, conforming more closely to gendered stereotypes regarding work norms.

The fact that we can only confidently identify 23 firms as engaging in discrimination against Black names when using a massive correspondence experiment reveals the difficulty of the signal extraction problem associated with estimating firm specific biases from application-level data. In future work, we will study how to use covariates to form improved forecasts that can be shared with the public. As described in Avivi et al. (2021), the firm-wide patterns documented here can potentially be used to design follow up correspondence experiments aimed at accurately measuring biases at particular jobs, information which may be of interest both to regulators and companies interested in monitoring their own behavior.

The variation in discrimination across employers documented here raises the question of whether bias at the most discriminatory firms can be reduced or eliminated by changes in organizational hiring practices. A large experimental psychology literature studying behavioral interventions designed to reduce prejudice has failed to produce a "silver bullet" treatment with proven effectiveness.[27] One of the strongest firm-level predictors of both racial and gender contact gaps found in our correspondence experiment is callback centralization. While centralizing interview decisions might serve to reduce discrimination, such changes may also simply postpone discrimination to a later stage of the hiring process. It is worth noting, however, that a growing industry purports to bolster diversity

[27] A recent review of this evidence by Paluck et al. (2020) concludes that "a fair assessment of our data on implicit prejudice reduction is that the evidence is thin. Together with the lack of evidence for diversity training, these studies do not justify the enthusiasm with which implicit prejudice reduction trainings have been received in the world over the past decade."

by automating portions of the recruiting process. Determining whether it is possible to improve recruiting practices in a way that promotes both equity and productivity remains an important and active area of research (Bergman, Li and Raymond, 2020; Raghavan et al., 2020).

References

Abaluck, Jason, Mauricio Caceres Bravo, Peter Hull, and Amanda Starc. 2021. "Mortality effects and choice across private health insurance plans." *The Quarterly Journal of Economics*, 136(3): 1557–1610.

Agan, Amanda, and Sonja Starr. 2018. "Ban the box, criminal records, and racial discrimination: A field experiment." *The Quarterly Journal of Economics*, 133(1): 191–235.

Agan, Amanda, and Sonja Starr. 2020. "Employer neighborhoods and racial discrimination." National Bureau of Economic Research.

Aigner, Dennis J, and Glen G Cain. 1977. "Statistical theories of discrimination in labor markets." *Ilr Review*, 30(2): 175–187.

Anatolyev, Stanislav, and Mikkel Sølvsten. 2020. "Testing Many Restrictions Under Heteroskedasticity." *arXiv preprint arXiv:2003.07320.*

Angrist, Joshua D, and Alan B Krueger. 1995. "Split-sample instrumental variables estimates of the return to schooling." *Journal of Business & Economic Statistics*, 13(2): 225–235.

Angrist, Joshua D., Peter D. Hull, Parag A. Pathak, and Christopher R. Walters. 2017. "Leveraging lotteries for school value-Added: Testing and estimation." *The Quarterly Journal of Economics*, 132(2): 871–919.

Arceo-Gomez, Eva O, and Raymundo M Campos-Vazquez. 2014. "Race and marriage in the labor market: A discrimination correspondence study in a developing country." *American Economic Review*, 104(5): 376–80.

Armstrong, Timothy. 2015. "Adaptive testing on a regression function at a point." *Annals of Statistics*, 43(5): 2086–2101.

Arnold, David, Will Dobbie, and Crystal S Yang. 2018. "Racial Bias in Bail Decisions." *The Quarterly Journal of Economics*, 133(4): 1885–1932.

Arnold, David, Will S Dobbie, and Peter Hull. 2020. "Measuring racial discrimination in bail decisions." National Bureau of Economic Research.

Avivi, Hadar, Patrick Kline, Evan Rose, and Christopher Walters. 2021. "Adaptive Correspondence Experiments." *AEA Papers and Proceedings*, 111: 43–48.

Baert, Stijn. 2018. "Hiring discrimination: an overview of (almost) all correspondence experiments since 2005." In *Audit studies: Behind the scenes with theory, method, and nuance.* 63–77. Springer.

Bai, Yuehao, Andres Santos, and Azeem M Shaikh. 2021. “A Two-Step Method for Testing Many Moment Inequalities.” *Journal of Business & Economic Statistics*, 1–33.

Banerjee, Rupa, Jeffrey G Reitz, and Phil Oreopoulos. 2018. “Do large employers treat racial minorities more fairly? An analysis of Canadian field experiment data.” *Canadian Public Policy*, 44(1): 1–12.

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

Becker, Gary S. 1993. “Nobel lecture: The economic way of looking at behavior.” *Journal of Political Economy*, 101(3): 385–409.

Benjamini, Yoav, and Yosef Hochberg. 1995. “Controlling the false discovery rate: a practical and powerful approach to multiple testing.” *Journal of the Royal statistical society: series B (Methodological)*, 57(1): 289–300.

Bergman, Peter, Danielle Li, and Lindsey Raymond. 2020. “Hiring as exploration.” *Available at SSRN 3630630.*

Bertrand, Marianne, and Esther Duflo. 2017. “Field experiments on discrimination.” In *Handbook of Field Experiments.* Vol. 1, , ed. Esther Duflo and Abhijit Banerjee. Elsevier.

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

Bertrand, Marianne, Sandra E Black, Sissel Jensen, and Adriana Lleras-Muney. 2019. “Breaking the glass ceiling? The effect of board quotas on female labour market outcomes in Norway.” *The Review of Economic Studies*, 86(1): 191–239.

Black, Dan A. 1995. “Discrimination in an equilibrium search model.” *Journal of labor Economics*, 13(2): 309–334.

Bohren, J Aislinn, Kareem Haggag, Alex Imas, and Devin G Pope. 2019. “Inaccurate statistical discrimination.” National Bureau of Economic Research.

Bostock v. Clayton County, Georgia. 590 U.S. 1-23 (2020). `https://www.supremecourt.gov/opinions/19pdf/17-1618_hfci.pdf`.

Bowlus, Audra J, and Zvi Eckstein. 2002. “Discrimination and skill differences in an equilibrium search model.” *International Economic Review*, 43(4): 1309–1345.

Canay, Ivan A, Magne Mogstad, and Jack Mountjoy. 2020. "On the use of outcome tests for detecting bias in decision making." National Bureau of Economic Research.

Charles, Kerwin Kofi, and Jonathan Guryan. 2008. "Prejudice and wages: an empirical assessment of Becker's The Economics of Discrimination." *Journal of Political Economy*, 116(5): 773–809.

Chetty, Raj, and Nathaniel Hendren. 2018. "The impacts of neighborhoods on intergenerational mobility I: Childhood exposure effects." *The Quarterly Journal of Economics*, 133(3): 1107–1162.

Chetty, Raj, John N Friedman, and Jonah E Rockoff. 2014. "Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood." *American economic review*, 104(9): 2633–79.

Crenshaw, Kimberle. 1989. "Demarginalizing the intersection of race and sex: a black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. Chicagounbound. uchicago. edu. 2018."

Crenshaw, Kimberle. 1990. "Mapping the margins: Intersectionality, identity politics, and violence against women of color." *Stanford Law Review*, 43: 1241.

Dahl, Gordon B, Andreas Kotsadam, and Dan-Olof Rooth. 2021. "Does integration change gender attitudes? The effect of randomly assigning women to traditionally male teams." *The Quarterly Journal of Economics*, 136(2): 987–1030.

Deming, David J. 2017. "The growing importance of social skills in the labor market." *The Quarterly Journal of Economics*, 132(4): 1593–1640.

Deming, David J, Noam Yuchtman, Amira Abulafi, Claudia Goldin, and Lawrence F Katz. 2016. "The value of postsecondary credentials in the labor market: An experimental study." *American Economic Review*, 106(3): 778–806.

Ding, Peng. 2017. "A paradox from randomization-based causal inference." *Statistical science*, 331–345.

Donoho, David, and Jiashun Jin. 2004. "Higher criticism for detecting sparse heterogeneous mixtures." *The Annals of Statistics*, 32(3): 962–994.

Efron, Bradley. 2016. "Empirical Bayes deconvolution estimates." *Biometrika*, 103(1): 1–20.

Efron, Bradley, and Robert Tibshirani. 1996. "Using specially designed exponential families for density estimation." *Annals of Statistics*, 24(6): 2431–2461.

Efron, Bradley, Robert Tibshirani, John D Storey, and Virginia Tusher. 2001. "Empirical Bayes analysis of a microarray experiment." *Journal of the American statistical association*, 96(456): 1151–1160.

Fang, Albert H, Andrew M Guess, and Macartan Humphreys. 2019. "Can the government deter discrimination? Evidence from a randomized intervention in New York City." *The Journal of Politics*, 81(1): 127–141.

Feigenberg, Benjamin, and Conrad Miller. 2021. "Would eliminating racial disparities in motor vehicle searches have efficiency costs?" *The Quarterly Journal of Economics.* qjab018.

Fryer Jr, Roland G, and Steven D Levitt. 2004. "The causes and consequences of distinctively black names." *The Quarterly Journal of Economics*, 119(3): 767–805.

Gaddis, S. Michael. 2017. "How Black Are Lakisha and Jamal? Racial perceptions from names used in correspondence audit studies." *Sociological Science*, 4(19): 469–489.

Goldin, Claudia. 2014. "A pollution theory of discrimination: male and female differences in occupations and earnings." In *Human capital in history: The American record.* 313–348. University of Chicago Press.

Gu, Jiaying, and Roger Koenker. 2020. "Invidious Comparisons: Ranking and Selection as Compound Decisions." *arXiv preprint arXiv:2012.12550.*

Holzer, Harry J, and Keith R Ihlanfeldt. 1998. "Customer discrimination and employment outcomes for minority workers." *The Quarterly Journal of Economics*, 113(3): 835–867.

Huang, Bert I. 2004. "The'Inexorable Zero'." *Harvard Law Review*, 117(4): 1215.

Hull, Peter. 2021. "What marginal outcome tests can tell us about racially biased decision-making." National Bureau of Economic Research.

Kane, Thomas J, and Douglas O Staiger. 2008. "Estimating teacher impacts on student achievement: An experimental evaluation." National Bureau of Economic Research.

Kline, Patrick M, and Christopher R Walters. 2021. "Reasonable doubt: Experimental detection of job-level employment discrimination." *Econometrica*, 89(2): 765–792.

Kline, Patrick, Raffaele Saggio, and Mikkel Sølvsten. 2020. "Leave-out estimation of variance components." *Econometrica*, 88(5): 1859–1898.

Leonard, Jonathan S, David I Levine, and Laura Giuliano. 2010. "Customer discrimination." *The Review of Economics and Statistics*, 92(3): 670–678.

Narasimhan, Balasubramanian, and Bradley Efron. 2020. "deconvolveR: A G-Modeling Program for Deconvolution and Empirical Bayes Estimation." *Journal of Statistical Software*, 94(1): 1–20.

Neumark, David, Ian Burn, and Patrick Button. 2018. "Is it harder for older workers to find jobs? New and improved evidence from a field experiment." *Journal of Political Economy*, 127(2): 922–970.

Nunley, John M, Adam Pugh, Nicholas Romero, and R Alan Seals. 2015. "Racial discrimination in the labor market for recent college graduates: Evidence from a field experiment." *The BE Journal of Economic Analysis & Policy*, 15(3): 1093–1125.

Office of Federal Contract Compliance Programs, U.S. Department of Labor v. Bank of America. ARB Case No. 13-099, ALJ Case No. 1997-OFC-016. (2016). `https://www.oalj.dol.gov/PUBLIC/ARB/DECISIONS/ARB_DECISIONS/OFC/13_099.OFCP.PDF`.

Office of Federal Contract Compliance Programs, U.S. Department of Labor v. Enterprise RAC Company of Baltimore, LLC. Case No.: 2016-OFC-00006. (2019). `https://www.oalj.dol.gov/DECISIONS/ALJ/OFC/2016/ENTERPRISE_RAC_COMPA_v_OFCCP_-_WASHINGTON_D_20160FC00006_(JUL_17_2019)_103111_CADEC_PD.PDF?_ga=2.224300827.1228285145.1624651131-670965852.1624651131`.

Onwuachi-Willig, Angela, and Mario L. Barnes. 2005. "By any other name: on being regarded as Black, and why Title VII should apply even if Lakisha and Jamal are white." *Wisconsin Law Review*, 1283.

Pager, Devah. 2016. "Are firms that discriminate more likely to go out of business?" *Sociological Science*, 3: 849–859.

Pager, Devah, Bart Bonikowski, and Bruce Western. 2009. "Discrimination in a low-wage labor market: A field experiment." *American sociological review*, 74(5): 777–799.

Paluck, Elizabeth Levy, Roni Porat, Chelsey S Clark, and Donald P Green. 2020. "Prejudice reduction: Progress and challenges." *Annual Review of Psychology*, 72.

Quillian, Lincoln, Devah Pager, Ole Hexel, and Arnfinn H Midtbøen. 2017. "Meta-analysis of field experiments shows no change in racial discrimination in hiring over time." *Proceedings of the National Academy of Sciences*, 114(41): 10870–10875.

Quillian, Lincoln, John J Lee, and Mariana Oliver. 2020. "Evidence from field experiments in hiring shows substantial additional racial discrimination after the callback." *Social Forces*, 99(2): 732–759.

Raghavan, Manish, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. "Mitigating bias in algorithmic hiring: Evaluating claims and practices." 469–481.

Rose, Evan K. 2020. "Who gets a second chance? Effectiveness and equity in supervision of criminal offenders." *The Quarterly Journal of Economics*, 136(2): 1199–1253.

Silverman, Bernard W. 1986. *Density estimation for statistics and data analysis.* Vol. 26, CRC press.

Stephens-Davidowitz, Seth. 2014. "The cost of racial animus on a black candidate: Evidence using Google search data." *Journal of Public Economics*, 118: 26–40.

Storey, John D. 2002. "A direct approach to false discovery rates." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3): 479–498.

Storey, John D. 2003. "The positive false discovery rate: a Bayesian interpretation and the q-value." *The Annals of Statistics*, 31(6): 2013–2035.

Storey, John D, Andrew J Bass, Alan Dabney, and David Robinson. 2015. "qvalue: Q-value estimation for false discovery rate control. R package." `https://github.com/StoreyLab/qvalue`.

Storey, John D, and Robert Tibshirani. 2003. "Statistical significance for genomewide studies." *Proceedings of the National Academy of Sciences*, 100(16): 9440–9445.

Teamsters v. United States. 431 U.S. 324 (1977). `https://supreme.justia.com/cases/federal/us/431/324/`.

Tilcsik, Andras. 2011. "Pride and prejudice: employment discrimination against openly gay men in the United States." *American Journal of Sociology*, 117(2): 586—-626.

U.S. EEOC. 1996. "Enforcement Guidance: Whether "testers" can file charges and litigate claims of employment discrimination." EEOC Notice No. N-915.002. `https://www.eeoc.gov/laws/guidance/enforcement-guidance-whether-testers-can-file-charges-and-litigate-claims-employment`.

U.S. EEOC. 2006*a*. "Directives Transmittal: Section 15 Race and Color Discrimination." No. 915.003. OLC Control No. EEOC-CVG-2006-1. `https://www.eeoc.gov/laws/guidance/section-15-race-and-color-discrimination`.

U.S. EEOC. 2006*b*. "Systemic Task Force Report To the Chair of the Equal Employment Opportunity Commission."

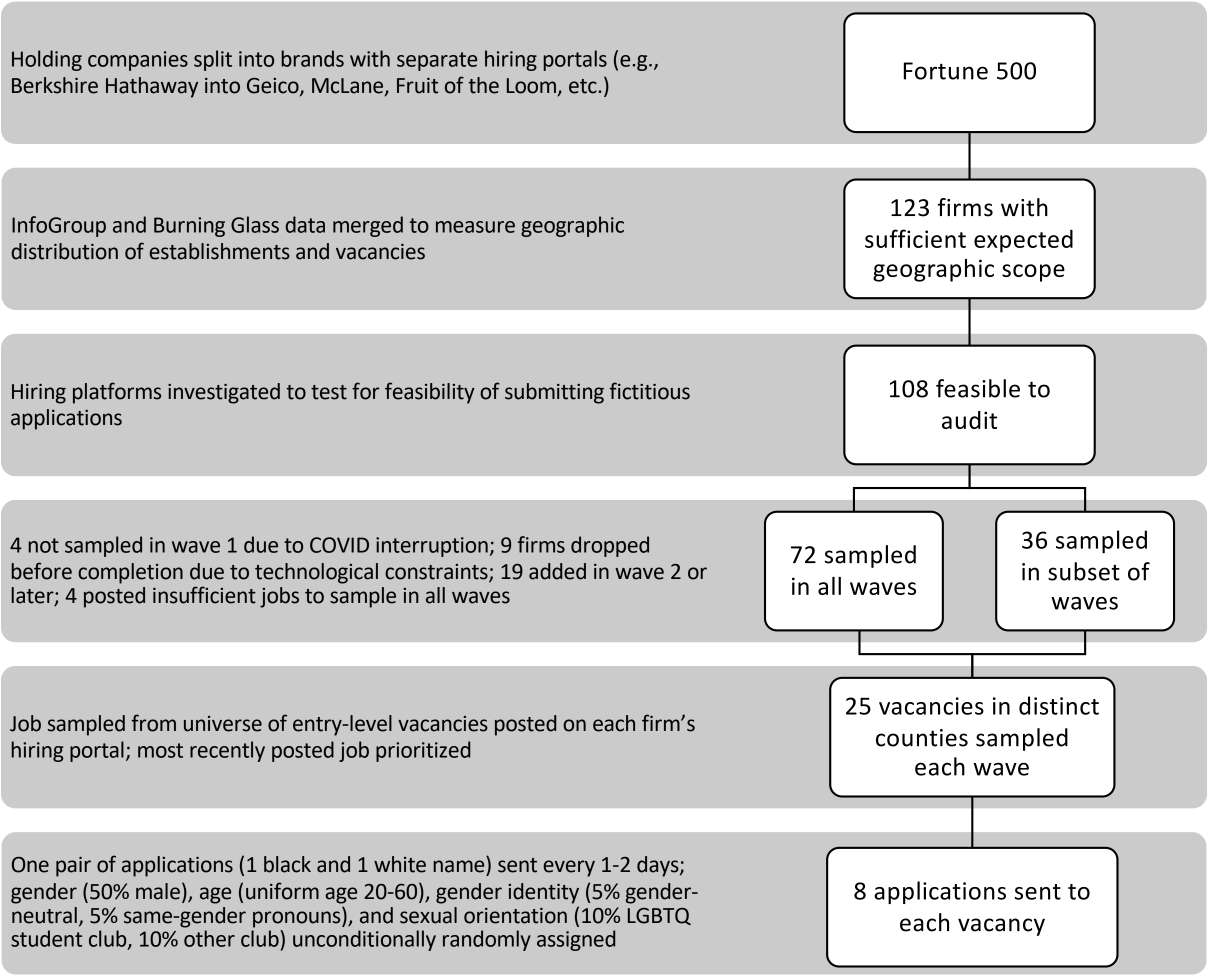
U.S. EEOC. 2019. "EEOC Sues Schuster for Sex Discrimination." *Press Release*, `https://www.eeoc.gov/newsroom/eeoc-sues-schuster-sex-discrimination`.

U.S. EEOC. 2020. "Sactacular Holdings to Pay $35,000 to Settle EEOC Sex Discrimination Lawsuit." *Press Release*, `https://www.eeoc.gov/newsroom/sactacular-holdings-pay-35000-settle-eeoc-sex-discrimination-lawsuit?utm_source=elinfonet`.

U.S. Equal Employment Opportunity Commission v. Target Corp. 460 F.3d 946 (7th Cir. 2006). `https://casetext.com/case/us-eeoc-v-target-corp`.

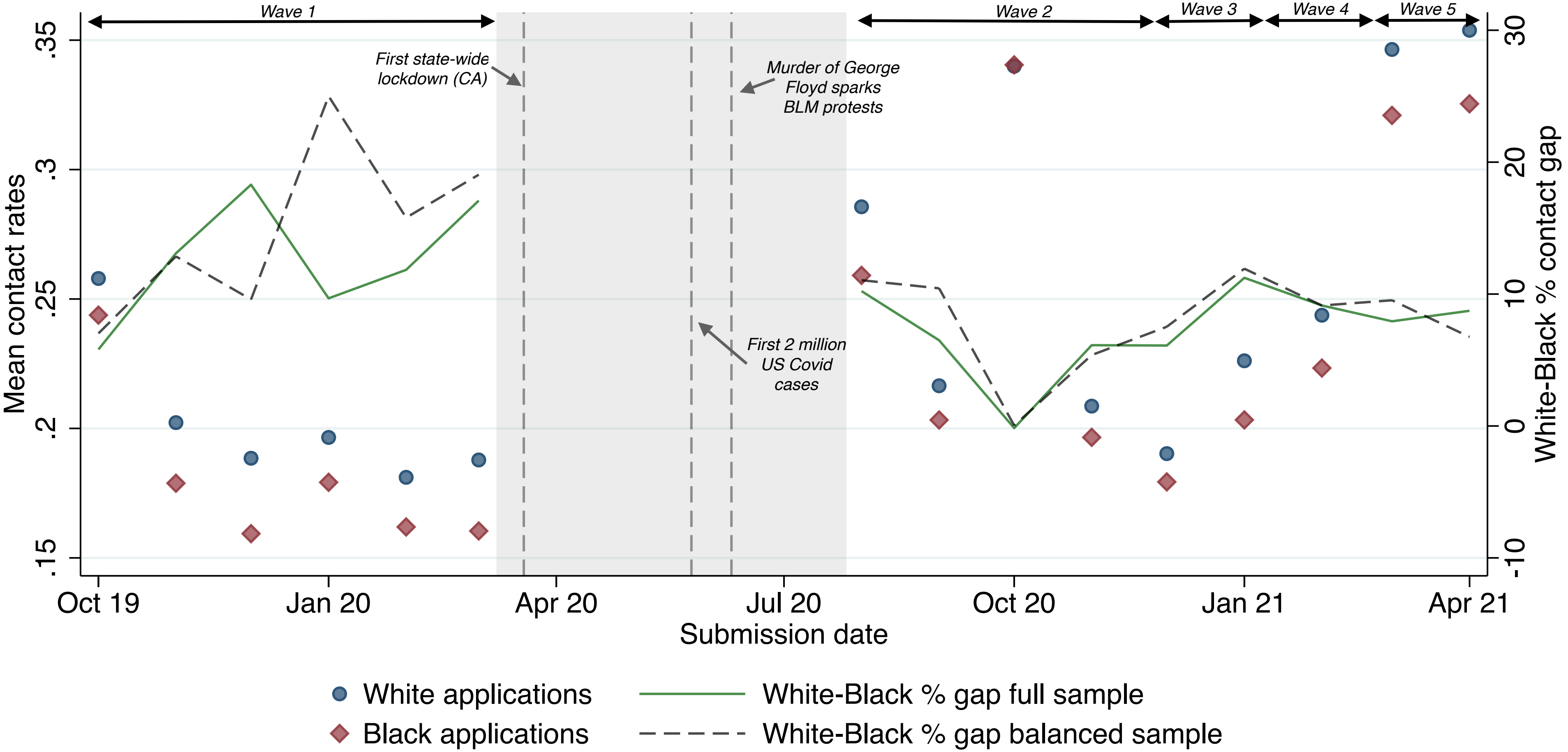
Figures

Figure 1: Overview of sampling strategy and experimental design


Holding companies split into brands with separate hiring portals (e.g., Berkshire Hathaway into Geico, McLane, Fruit of the Loom, etc.)
Fortune 500
InfoGroup and Burning Glass data merged to measure geographic distribution of establishments and vacancies
123 firms with sufficient expected geographic scope
Hiring platforms investigated to test for feasibility of submitting fictitious applications
108 feasible to audit
4 not sampled in wave 1 due to COVID interruption; 9 firms dropped before completion due to technological constraints; 19 added in wave 2 or later; 4 posted insufficient jobs to sample in all waves
72 sampled in all waves
36 sampled in subset of waves
Job sampled from universe of entry-level vacancies posted on each firm's hiring portal; most recently posted job prioritized
25 vacancies in distinct counties sampled each wave
One pair of applications (1 black and 1 white name) sent every 1-2 days; gender (50% male), age (uniform age 20-60), gender identity (5% gender-neutral, 5% same-gender pronouns), and sexual orientation (10% LGBTQ student club, 10% other club) unconditionally randomly assigned
8 applications sent to each vacancy

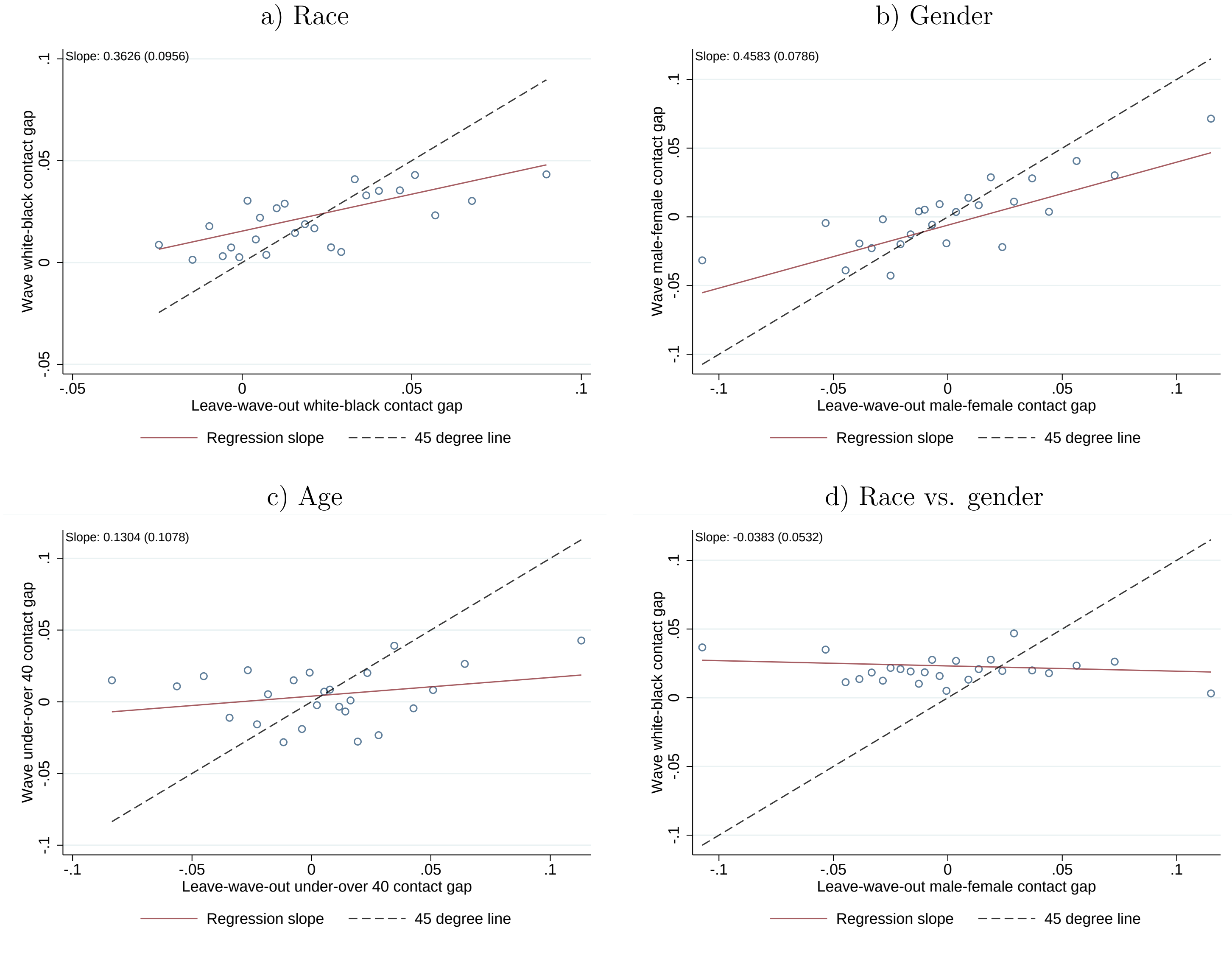
Notes: This figure explains the sampling strategy and design for the experiment. Gender identity and sexual orientation attributes were assigned starting in wave 2 after the U.S. Supreme Court ruling in Bostock v. Clayton County, Georgia.

Figure 2: Mean contact rates and racial contact gaps by date



Notes: This figure plots mean contact rates for black and white applications by month and year of submission. The gray region corresponds to the period when the experiment was paused due to Covid-19 related shut-downs. The green and dotted black lines plot the white-Black contact rate gap as a percentage of the mean Black contact rate for the full and balanced samples. An F-test fails to reject that the white/Black percentage point difference in contact rates is the same in all waves of the experiment ($F = 0.85$, $p = 0.50$).

Figure 3: Stability of firm contact gaps across waves



a) Race

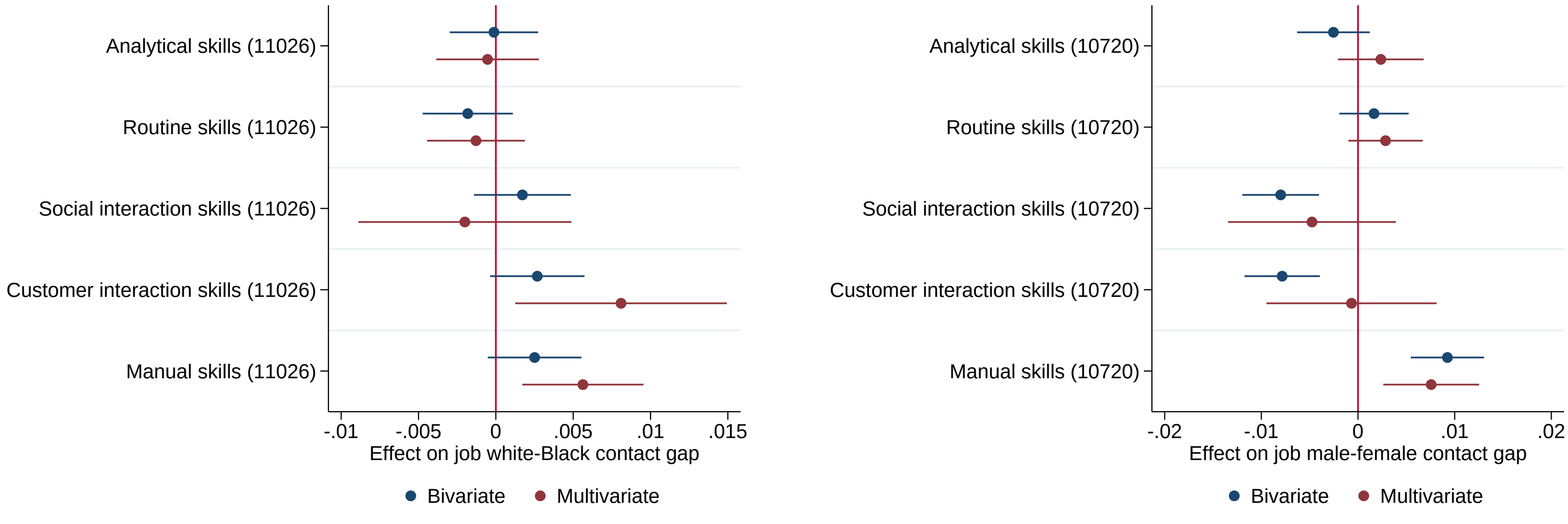
b) Gender

c) Age

d) Race vs. gender

Notes: This figure presents binned scatter plots of firm-specific wave-average contact gaps vs. leave-wave-out firm-specific average contact gaps. Panel (a) reports results for the white/Black difference in contact rates. Panel (b) shows results for the male/female difference in contact rates. Panel (c) displays results for the difference between contact rates for applicants under and over age 40. Panel (d) plots the correlation between race and gender contact gaps. The points are means of the dependent and independent variables within vigintiles of the independent variable. The dotted line has a slope of 1 and passes through the origin. The red line corresponds to the regression slope reported on the figure, with firm-clustered standard errors reported in parentheses. All firms present in at least 2 waves are included.

Figure 4: Relationships between contact gaps and job task content



a) Race

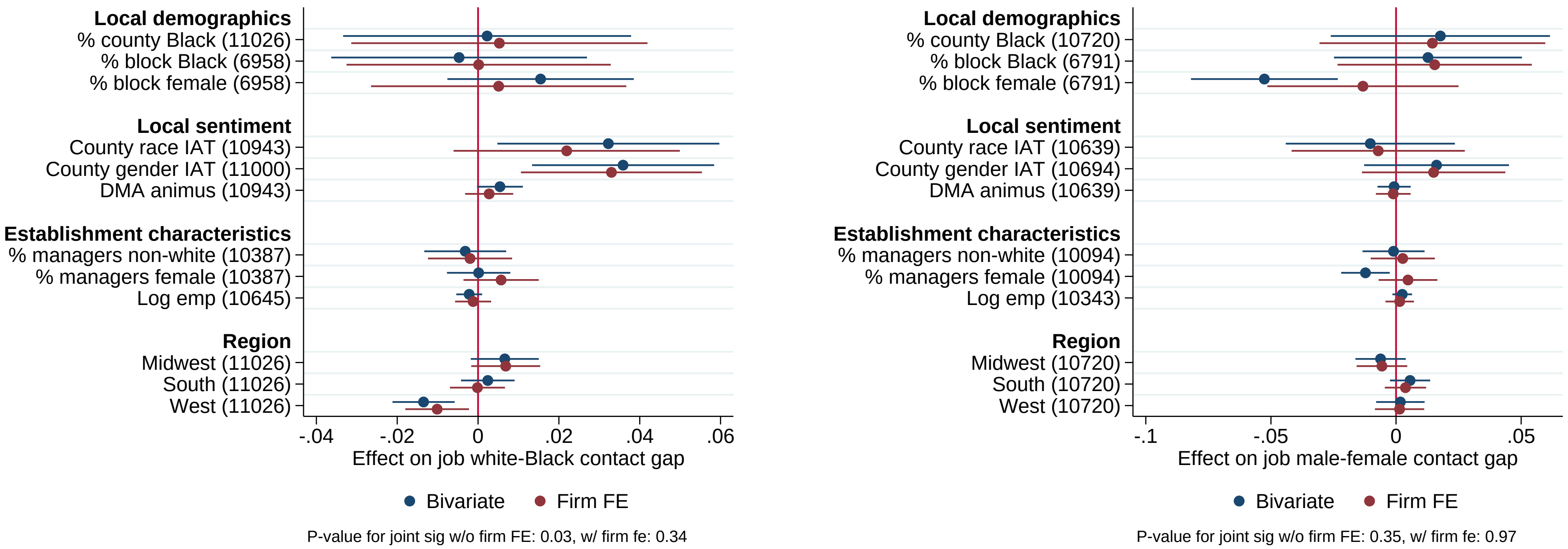
P-value for joint sig w/o firm FE: 0.01, w/ firm fe: 0.20

b) Gender

P-value for joint sig w/o firm FE: 0.00, w/ firm fe: 0.01

Notes: This figure plots the relationship between O*Net measures of job-level task content and contact gaps for race and gender. Each relationship is estimated with a linear regression with job-level contact gaps as the outcome. All jobs with defined contact gaps for each attribute are included. The number of jobs in each regression is in parentheses. Task measures are normalized to have standard deviation one in sample. "Bivariate" points plot coefficients from regressions of contact gaps on the covariate alone. "Multivariate" points plot effects when all covariates are included simultaneously. Bars indicate 95% confidence intervals based on robust standard errors. Appendix C provides a complete description of task definitions and sources.

Figure 5: Relationships between contact gaps and establishment characteristics

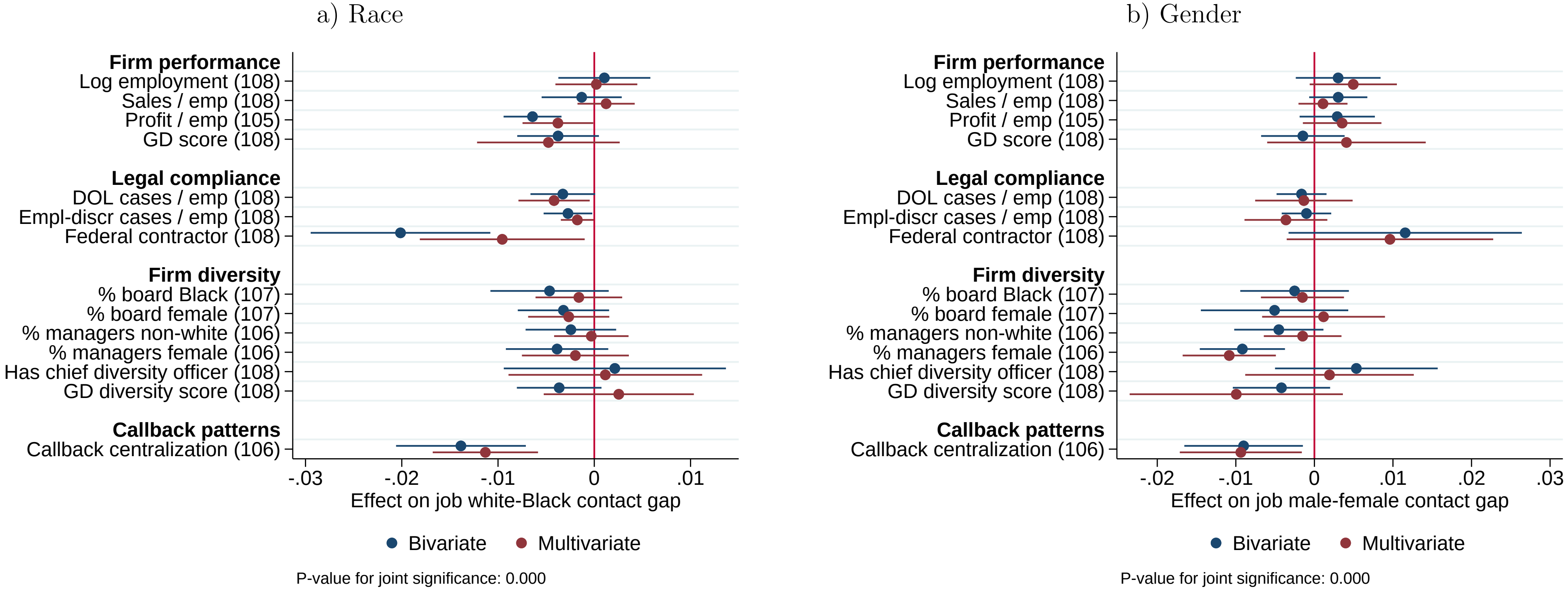


a) Race

b) Gender

Notes: This figure plots the relationship between establishment-level covariates and contact gaps for race and gender. Each relationship is estimated with a linear regression with job-level contact gaps as the outcome. All jobs with defined contact gaps for each attribute and matched to the listed covariate are included. The number of jobs in each regression is in parentheses. Covariates are standardized to be mean zero, standard deviation 1 in sample. “Bivariate” points plot coefficients from regressions of contact gaps on the covariate alone. “Firm FE” points include firm fixed effects. Bars indicate 95% confidence intervals based on robust standard errors. The omitted region category is the Northeast. Appendix C provides a complete description of covariate definitions and sources.

Figure 6: Relationships between contact gaps and firm characteristics

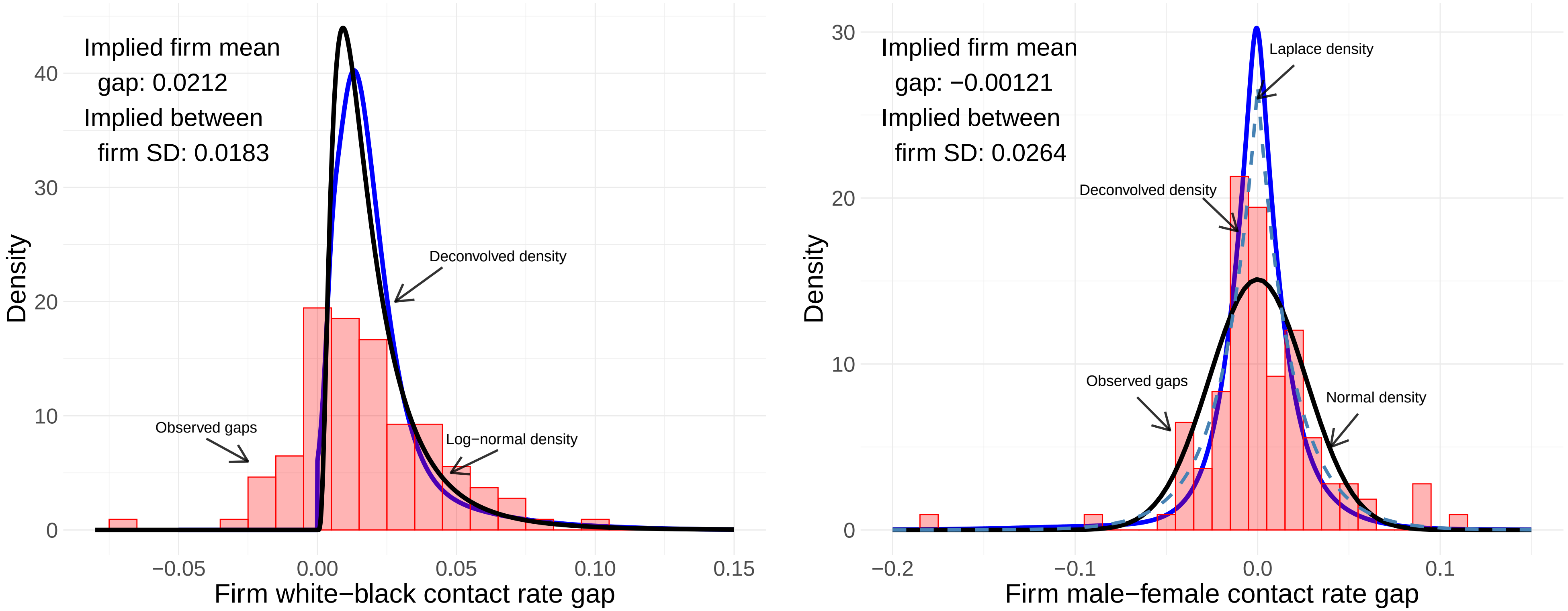


Notes: This figure plots the relationship between firm-level covariates and contact gaps for race and gender. Each relationship is estimated by a linear regression with job-level contact gaps as the outcome, except for callback centralization which is estimated via split sample IV to account for any mechanical correlation with the outcome. The number of firms in each regression is in parentheses. Covariates are standardized to be mean zero, standard deviation 1 in sample. "Bivariate" points plot coefficients from regressions of contact gaps on the covariate alone. "Multivariate" points plot effects when all covariates are included simultaneously, with callback centralization measured in half of the randomly split sample instrumented using its value in the other half and all other covariates included as exogenous regressors. Bars indicate 95% confidence intervals based on standard errors clustered at the firm level. Appendix C provides a complete description of covariate definitions and sources.

Figure 7: Deconvolution estimates of firm-level discrimination distributions

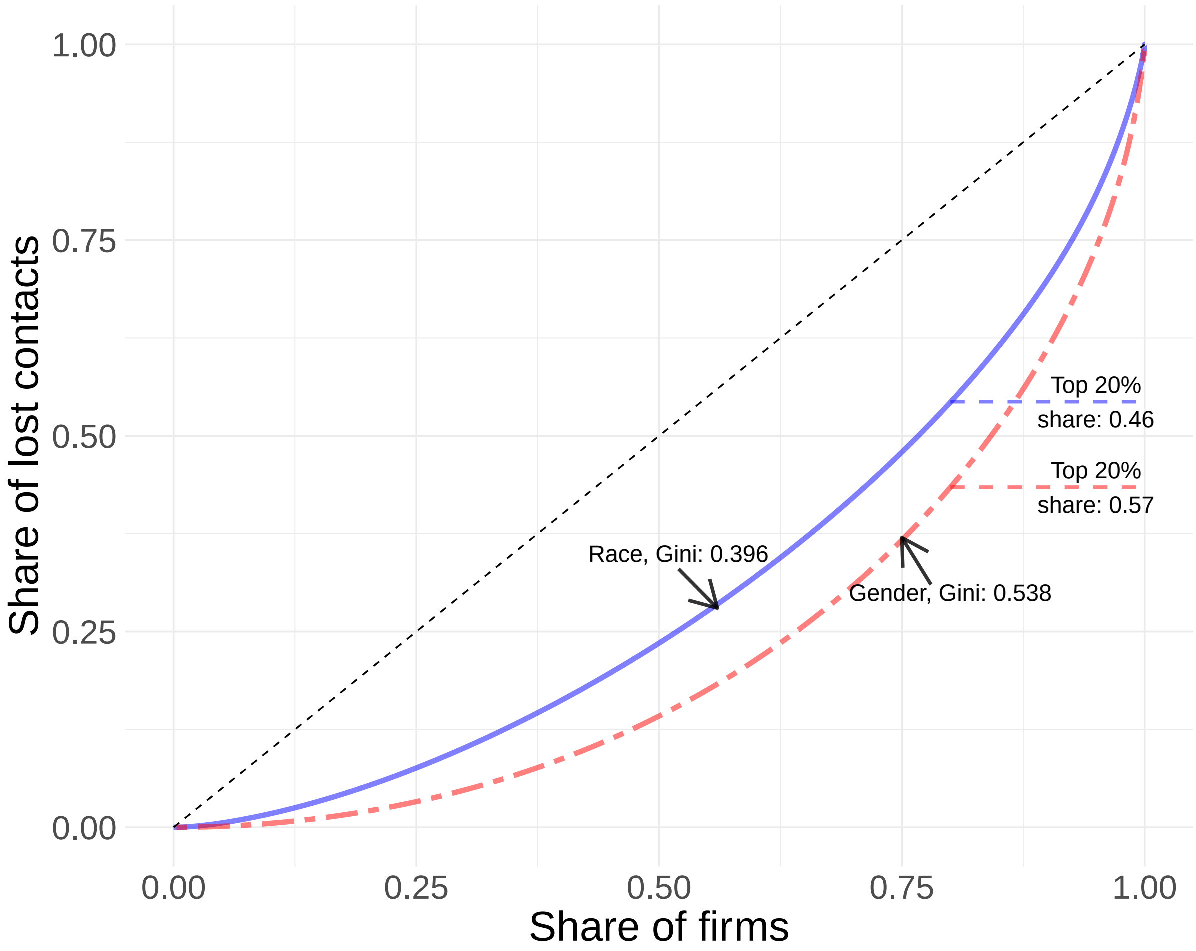
a) Race

b) Gender



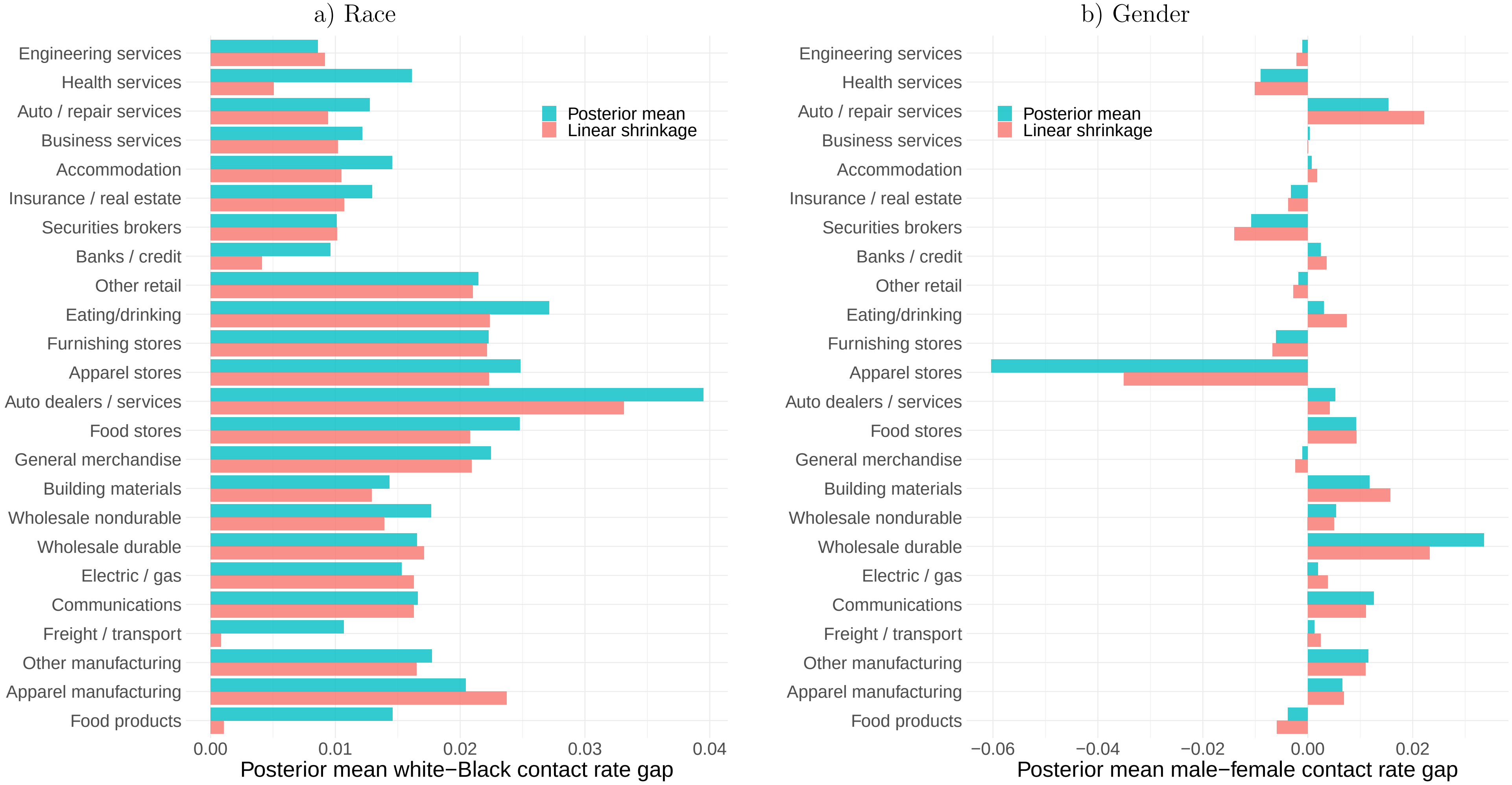
Notes: This figure presents non-parametric estimates of the distribution of firm-specific contact gaps. Panel (a) presents estimates for white-Black contact rate differences, and panel (b) presents estimates for male-female differences. Red histograms show the distribution of estimated firm contact gaps. Blue lines shows estimates of population contact gap distributions. The population distributions are estimated by applying the deconvolveR package (Narasimhan and Efron, 2020) to firm-specific z-score estimates, then numerically integrating over the distribution of log standard errors using a Gaussian kernel and the Silverman (1986) bandwidth to recover the distribution of contact gaps. The penalization parameter in the deconvolution step is calibrated so that the resulting distribution matches the corresponding bias-corrected variance estimate from Table 4. In panel (a), the density of population z-scores is constrained to be weakly positive. Parametric comparison distributions are calibrated to have means and variances matching those of the deconvolved distributions.

Figure 8: Discrimination Lorenz curves



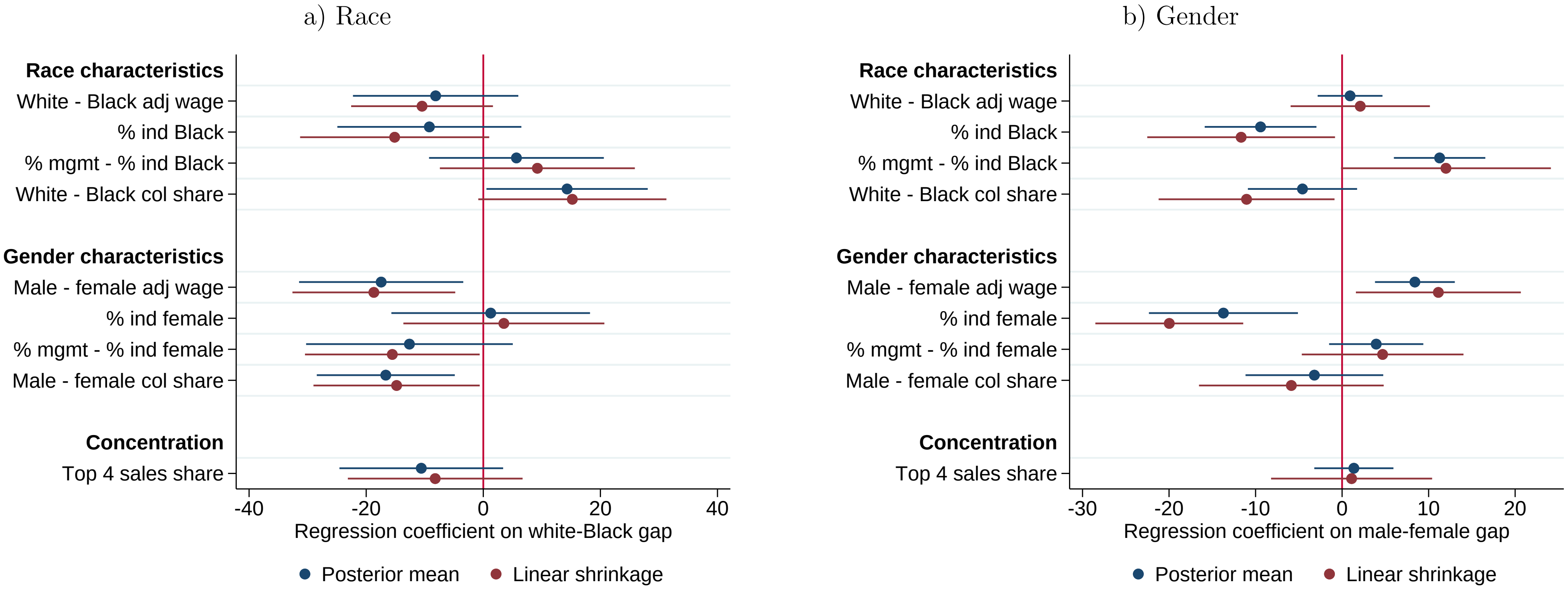
Notes: This figure displays Lorenz curves implied by the non-parametric deconvolution estimates of race and gender contact gap distributions in Figure 7. The solid blue curve is the Lorenz curve for the white/Black contact gap, and the dashed red curve is the Lorenz curve for the absolute value of the male/female contact gap. The Lorenz curve reports the share of lost contacts in the experiment attributable to firms below each contact gap percentile. The share of lost contacts equals the sum of contact gaps at firms below a particular contact gap percentile as a share of the sum of contact gaps across all firms. The dashed line is the 45 degree line. The labels for each curve also report Gini coefficients, equal to 1 minus twice the area under each curve.

Figure 9: Posterior means by industry



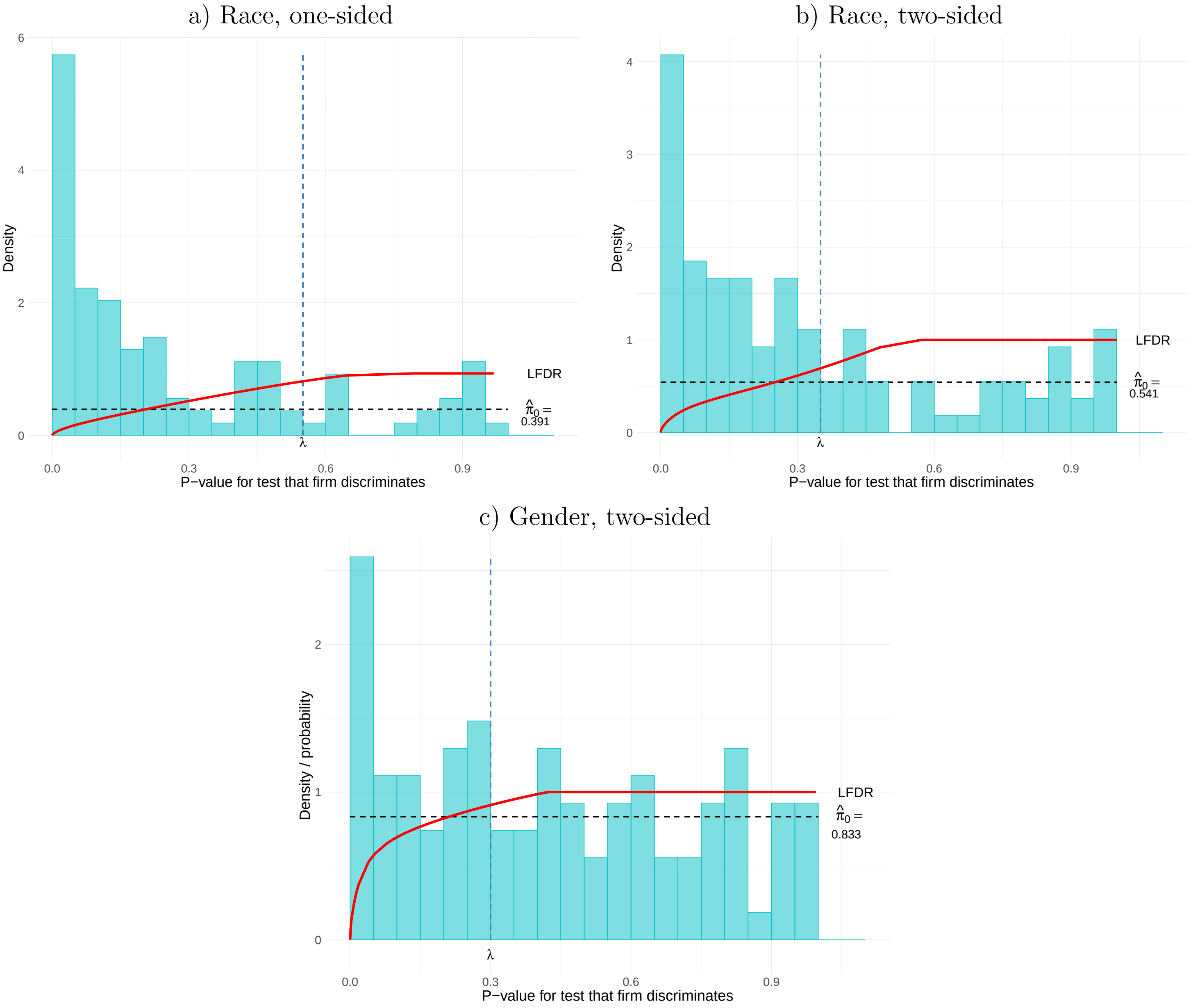
Notes: This figure presents industry-level averages of posterior mean contact gaps. The pink bars show average posteriors using deconvolved estimates of the population contact gaps as priors. The teal bars show averages of estimates shrunk linearly towards the grand mean with weights given by the signal to noise ratio $\hat{\theta}/(s_f^2 + \hat{\theta})$.

Figure 10: Industry correlates of contact gaps



Notes: This figure presents regressions of industry characteristics on posterior mean contact gaps for race and gender. Points labeled "posterior means" show coefficients on posterior gaps formed using the distributions in Figure 7 as priors. Points labeled "linear shrinkage" show coefficients on gaps shrunk linearly towards the grand mean with weights given by the signal to noise ratio $\hat{\theta}/(s_f^2 + \hat{\theta})$. Outcomes are normalized to be mean zero, standard deviation 1 in sample. Appendix C provides a complete description of covariate definitions and sources.

Figure 11: *P*-value distributions and local false discovery rates



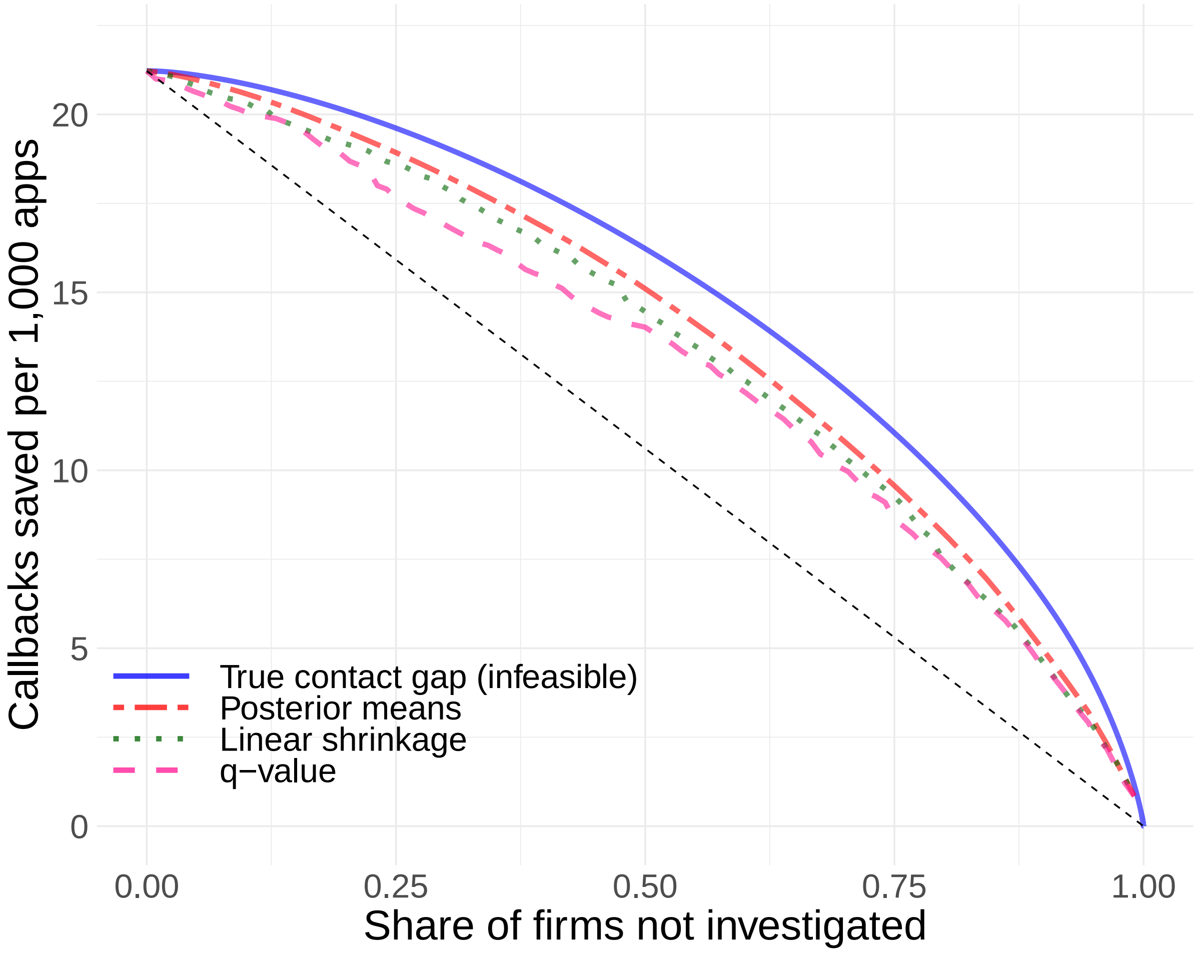
a) Race, one-sided

b) Race, two-sided

c) Gender, two-sided

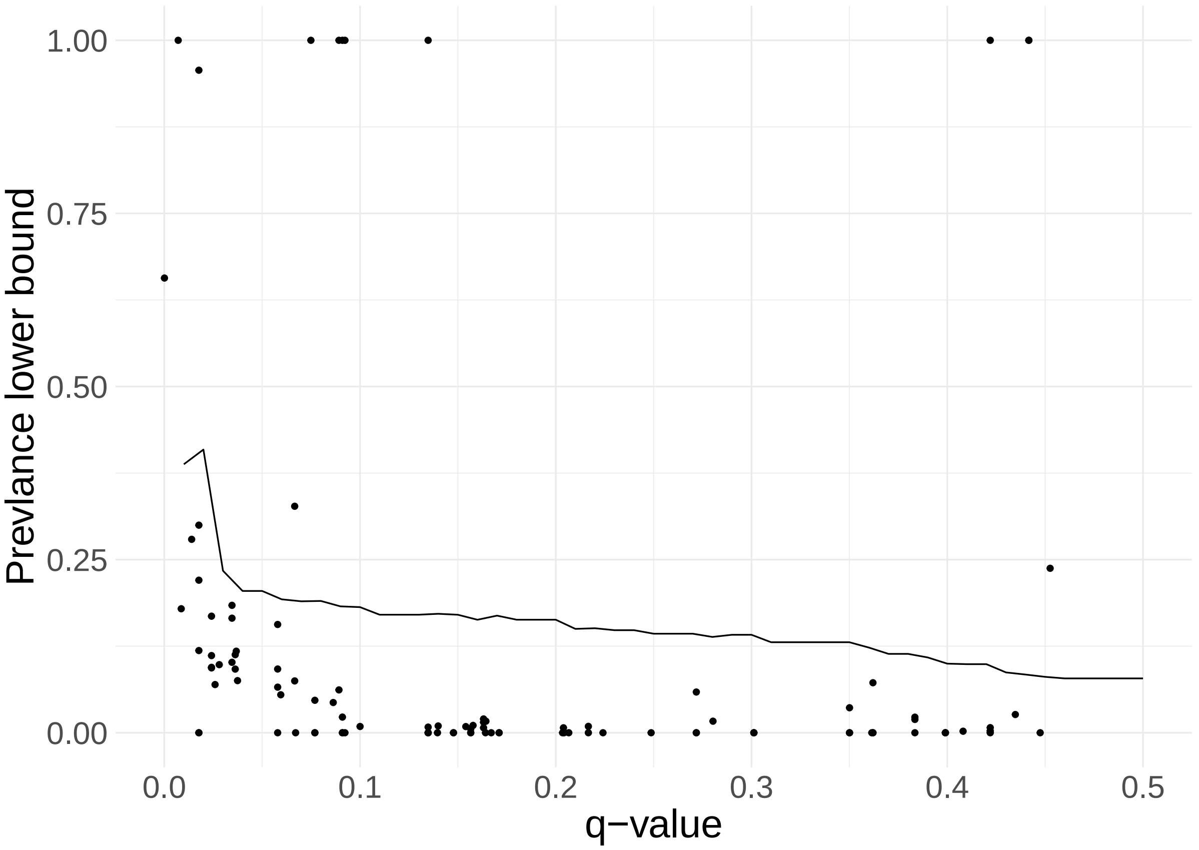
Notes: This figure plots distributions of *p*-values from firm-specific tests of the null hypothesis of no discrimination. Panel (a) shows results for one-sided tests of no discrimination against Black applicants, and panel (b) displays results for two-sided tests of equal contact rates for Black and white applicants. Panel (c) shows results for two-sided tests of equal contact rates for male and female applicants. Dotted black lines show estimated upper bounds on π_0, the share of non-discriminating firms. Red lines trace local false discovery rates. *P*-values comes from paired *t*-tests applied to job-level contact rate gaps for each firm.

Figure 12: Detection tradeoffs



Notes: This figure illustrates the expected number of contacts per thousand Black applications sent that would be saved if discrimination were eliminated at all firms below a ranking threshold. We consider four rankings: infeasible ranking by true contact gaps (Δ_f), ranking by posterior means ($\bar{\Delta}_f$), ranking by linear shrinkage estimates ($\tilde{\Delta}_f$), and ranking by q-values ($\hat{q}_f$). The dashed black line shows the results of ranking firms randomly.

Figure 13: Job-level prevalence of racial discrimination



Notes: This figure shows estimated lower bounds on the prevalence of job-level racial discrimination within firms. Each point depicts a firm's estimated lower bound prevalence, computed according to the formula $(\hat{\Delta}_f^2 - s_f^2)/(\hat{\sigma}_f^2 + \hat{\Delta}_f^2 - s_f^2)$, where $\hat{\sigma}_f^2$ is the job level covariance between contact gaps arising in the first four and last four applications. Firm-specific bound estimates have been constrained to fall in the unit interval. The black line plots prevalence bounds computed by pooling jobs from all firms with q-values less than the threshold depicted on the horizontal axis.

Tables

Table 1: Summary statistics

	A. All firms			B. Balanced sample		
	White	Black	Difference	White	Black	Difference
Resume characteristics						
Female	0.499	0.499	-0.001	0.500	0.498	0.003
Over 40	0.535	0.535	0.000	0.534	0.533	0.002
LGBTQ club member	0.081	0.082	-0.001	0.079	0.080	-0.001
Academic club	0.040	0.042	-0.002	0.039	0.042	-0.003*
Political club	0.042	0.042	0.001	0.042	0.041	0.001
Gender-neutral pronouns	0.041	0.041	-0.001	0.040	0.040	0.000
Same-gender pronouns	0.043	0.042	0.001	0.042	0.041	0.001
Associate degree	0.476	0.485	-0.009**	0.478	0.485	-0.006*
Geographic distribution						
Northeast	0.150	0.150	-0.000	0.152	0.152	-0.000
Midwest	0.220	0.220	0.000	0.221	0.221	0.000
South	0.416	0.416	-0.000	0.423	0.423	-0.000
West	0.214	0.214	0.000	0.204	0.204	-0.000
Wave distribution						
Wave 1	0.174	0.174	0.000	0.189	0.189	0.000
Wave 2	0.206	0.206	0.000	0.210	0.210	0.000
Wave 3	0.215	0.215	-0.000	0.204	0.204	-0.000
Wave 4	0.205	0.205	-0.000	0.198	0.198	-0.000
Wave 5	0.200	0.200	-0.000	0.199	0.199	-0.000
Contact rates						
Any contact in 30 days	0.251	0.230	0.020***	0.256	0.234	0.022***
Voicemail	0.178	0.159	0.019***	0.185	0.166	0.019***
Email	0.040	0.039	0.002	0.043	0.042	0.002
Text	0.033	0.032	0.000	0.028	0.027	0.001
Any contact in 14 days	0.217	0.199	0.017***	0.222	0.203	0.019***
Any contact in 15-30 days	0.034	0.031	0.003***	0.034	0.031	0.003**
N applications	41837	41806	83643	32703	32665	65368
N jobs			11114			8667
N firms			108			72
1/2/3/4/5 waves			3/4/15/16/72			

Notes: This table presents summary statistics for the full analysis sample and balanced sample of firms sent applications in all five waves of the experiment. “White” refers to resumes with distinctively white names; “Black” refers to resumes with distinctively Black names. LGBTQ club membership and gender-neutral pronouns were introduced in wave 2. Stars indicate significant differences from zero at the following levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 2: Effects of resume characteristics on contact rates

	A. All firms		B. Balanced sample	
	(1)	(2)	(3)	(4)
	LPM	Logit	LPM	Logit
Black	-0.0205***	-0.115***	-0.0222***	-0.123***
	(0.00169)	(0.00949)	(0.00193)	(0.0107)
Female	0.000184	0.000760	-0.000249	-0.00166
	(0.00300)	(0.0168)	(0.00341)	(0.0189)
Over 40	-0.00587**	-0.0332**	-0.00472	-0.0265
	(0.00299)	(0.0167)	(0.00341)	(0.0189)
Political club	-0.00180	-0.00985	-0.00316	-0.0172
	(0.00742)	(0.0406)	(0.00848)	(0.0458)
Academic club	0.00976	0.0520	0.00550	0.0283
	(0.00764)	(0.0407)	(0.00870)	(0.0461)
LGBTQ club	-0.00513	-0.0287	-0.0000	-0.0007
	(0.00545)	(0.0302)	(0.00637)	(0.0342)
Same-gender pronouns	-0.0139*	-0.0765*	-0.0126	-0.0677
	(0.00735)	(0.0412)	(0.00848)	(0.0466)
Gender-neutral pronouns	-0.0104	-0.0572	-0.0174**	-0.0946**
	(0.00755)	(0.0421)	(0.00857)	(0.0477)
Associate degree	0.00119	0.00665	0.00254	0.0139
	(0.00303)	(0.0170)	(0.00345)	(0.0191)
Midwest	0.0631***	0.323***	0.0454***	0.230***
	(0.0120)	(0.0622)	(0.0136)	(0.0692)
South	-0.0297***	-0.170***	-0.0396***	-0.221***
	(0.0103)	(0.0577)	(0.0117)	(0.0638)
West	-0.0266**	-0.153**	-0.0386***	-0.216***
	(0.0114)	(0.0650)	(0.0131)	(0.0729)
Wave 2	0.0535***	0.318***	0.0510***	0.302***
	(0.0106)	(0.0633)	(0.0116)	(0.0691)
Wave 3	0.0102	0.0624	0.0167	0.102
	(0.0101)	(0.0650)	(0.0115)	(0.0722)
Wave 4	0.0393***	0.238***	0.0416***	0.249***
	(0.0105)	(0.0640)	(0.0118)	(0.0709)
Wave 5	0.151***	0.798***	0.162***	0.842***
	(0.0113)	(0.0614)	(0.0127)	(0.0674)
Constant	0.207***	-1.358***	0.219***	-1.292***
	(0.0113)	(0.0666)	(0.0127)	(0.0728)
N applications	83643	83643	65368	65368

Notes: This table presents the effects of randomized protected applicant characteristics on the probability of employer contact within 30 days. Panel (a) includes all firms, while panel (b) includes the balanced sample of firms sent applications in every wave of the experiment. Columns 1 and 3 are linear probability models. Columns 2 and 4 are logistic regressions. Standard errors in parentheses are clustered at the job level. Stars indicate statistical significance at the following levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 3: Effects of race interacted with resume characteristics

	A. OLS			B. Logit		
	(1)	(2)	(3)	(4)	(5)	(6)
	White	Black	Difference	White	Black	Difference
Female	0.00716*	-0.00694*	0.0141**	0.0388*	-0.0398*	0.0786**
	(0.00423)	(0.00412)	(0.00579)	(0.0229)	(0.0236)	(0.0322)
Over 40	-0.0104**	-0.00125	-0.00915	-0.0562**	-0.00711	-0.0491
	(0.00428)	(0.00413)	(0.00590)	(0.0231)	(0.0236)	(0.0328)
Political club	-0.00207	-0.00229	0.000220	-0.0109	-0.0126	0.00171
	(0.0107)	(0.0105)	(0.0150)	(0.0562)	(0.0587)	(0.0815)
Academic club	0.00341	0.0147	-0.0113	0.0173	0.0806	-0.0633
	(0.0111)	(0.0107)	(0.0155)	(0.0576)	(0.0574)	(0.0817)
LGBTQ club	-0.0165**	0.00631	-0.0228**	-0.0889**	0.0349	-0.124**
	(0.00787)	(0.00763)	(0.0110)	(0.0431)	(0.0419)	(0.0601)
Same-gender pronouns	-0.00971	-0.0165	0.00681	-0.0515	-0.0934	0.0420
	(0.0106)	(0.0101)	(0.0146)	(0.0571)	(0.0587)	(0.0816)
Gender-neutral pronouns	-0.0106	-0.0103	-0.000279	-0.0564	-0.0578	0.00138
	(0.0108)	(0.0105)	(0.0150)	(0.0581)	(0.0598)	(0.0830)
Associates degree	0.00573	-0.00152	0.00724	0.0309	-0.00869	0.0396
	(0.00431)	(0.00412)	(0.00584)	(0.0233)	(0.0236)	(0.0325)
Constant	0.201***	0.185***	0.0160***	-1.377***	-1.485***	0.108***
	(0.00848)	(0.00820)	(0.00621)	(0.0514)	(0.0538)	(0.0366)
N applications	41837	41806	83643	41837	41806	83643
χ^2 stat for joint significance			16.55			16.45
p-value			0.0351			0.0363

Notes: This table presents the effects of race interacted with other resume characteristics. Columns 1 and 3 show estimates of models for employer contact among white applicants, columns 2 and 4 display estimates for Black applicants, and columns 3 and 6 show differences in coefficients between white and Black applicants. Panel (a) uses linear probability models, while panel (b) uses logistic regression. All models control for wave indicators. χ^2 statistics and joint p-values come from tests that all differences in reported coefficients other than the constant term are zero. Standard errors in parentheses are clustered at the job level. Stars indicate statistical significance at the following levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 4: Firm-level heterogeneity in discrimination

	(1) χ^2 test of heterogeneity	(2) p-value for no discrim against:	Contact gap SD (3) Bias-corrected	 (4) Cross-Wave	 (5) Cross-State
Race	276.5	W: 1.00	0.0185	0.0168	0.0178
	[0.000]	B: 0.00	(0.0033)	(0.0034)	(0.0035)
Gender	205.2	M: 0.00	0.0267	0.0287	0.0269
	[0.000]	F: 0.05	(0.0041)	(0.0037)	(0.0041)
Over 40	144.6	Y: 0.21	0.0103	0.0044	0.0086
	[0.011]	O: 0.03	(0.0053)	(0.0098)	(0.0059)

Notes: This table presents estimated standard deviations of firm-level contact rate gaps and tests for heterogeneity in gaps. Column 1 displays χ^2 test statistics and associated p-values from tests of the null hypothesis of no heterogeneity in discrimination. The test statistic is $\sum_f (\hat{\Delta}_f - \bar{\Delta})^2/s_f^2$, where $\hat{\Delta}_f$ is the contact cap estimate for firm f, s_f is the estimate's standard error, and $\bar{\Delta}$ is the equally-weighted average of contact gaps. Column 2 presents tests for one-sided discrimination against white (W), black (B), male (M), female (F), aged under 40 (Y), and over 40 (O) applications using the methodology in Bai, Santos and Shaikh (2021). Column 3 reports estimates of the standard deviation of average contact gaps across firms calculated using firm-specific standard errors to correct for bias due to sampling variation in $\hat{\Delta}_f$. Columns 4 and 5 report cross-wave and cross-state estimates based on covariances between firm-by-wave and firm-by-state contact gaps. Details on these estimators appear in Appendix D. Standard errors for all variance estimators are produced by job-clustered weighted bootstrap. Estimates include all 108 firms.

Table 5: Firm contact gap heterogeneity in levels, log odds, and log proportions

	LPM		Logit		Poisson	
	(1) Intercept	(2) Slope	(3) Intercept	(4) Slope	(5) Intercept	(6) Slope
Mean	0.2547	-0.0187	-1.2715	-0.1102	-1.6046	-0.0853
	(0.0035)	(0.0018)	(0.0263)	(0.0142)	(0.0222)	(0.0123)
Std. dev.	0.1607	0.0186	0.9755	0.1155	0.7047	0.0837
	(0.0035)	(0.0036)	(0.0366)	(0.0361)	(0.0368)	(0.0370)
Corr. w/own slope	-0.4010	1.000	0.0519	1.000	0.0685	1.000
	(0.1123)	-	(0.1855)	-	(0.2360)	-
Corr. w/LPM slope	-0.4010	1.000	-0.4274	0.8944	-0.5045	0.8075
	(0.1123)	-	(0.1093)	(0.0953)	(0.1155)	(0.1448)
Number of firms	103		103		103	

Notes: This table reports estimated means, standard deviations, and correlations of firm-specific intercept and Black slope coefficients from models for employer contact. Columns 1-2 show results from linear probability models (LPMs; levels), columns 3-4 display results from logit models (log odds), and columns 5-6 show results from Poisson regression models (log proportions). Means are averages of firm-specific coefficients. Standard deviations are calculated by subtracting the average squared job-clustered standard error from the sample variance of parameter estimates, then taking the square root. Correlations are computed by subtracting the average job-clustered sampling covariance from the sample covariance of parameter estimates, then dividing by the product of estimated standard deviations. The analysis is restricted to the 103 firms with callback rates above 3 percent. Standard errors (computed by job-clustered weighted bootstrap) in parentheses.

Table 6: Heterogeneity across alternative job groupings

	(1) Race	(2) Gender	(3) Over 40
State	0.0076	-	-
	(0.0029)		
	[0.038]	[0.668]	[0.583]
Industry	0.0141	0.0190	0.0048
	(0.0022)	(0.0030)	(0.0040)
	[0.000]	[0.000]	[0.112]
Job title	0.0135	0.0111	0.0033
SOC3 code	(0.0022)	(0.0039)	(0.0071)
	[0.000]	[0.007]	[0.527]
Hiring platform	0.0059	0.0024	0.0024
intermediary	(0.0023)	(0.0071)	(0.0059)
	[0.008]	[0.049]	[0.212]

Notes: This table presents estimates of heterogeneity in average contact rate gaps across states, industries, job titles, and hiring platform intermediaries, along with the results of tests for no heterogeneity across each set of groups. Estimates are standard deviations of group-level contact rate gaps, computed using the same bias-corrected estimator employed in column 1 of Table 4. Group variance components are computed weighting jobs in inverse proportion to the number of jobs sampled from each job's parent firm, so that groupings that nest firms are weighted by the number of firms in each group. Standard errors, produced by job-clustered weighted bootstrap, are reported in parentheses. Dashes indicate negative variance estimates and hence undefined estimated standard deviations. P-values from χ^2 tests of no heterogeneity in group-level contact rates are reported in square brackets. The first panel groups jobs by state, with 51 states (including D.C.) represented in the experiment. The second panel groups firms by the 24 two-digit SIC codes in the data. The third panel groups by the 47 three-digit SOC3 codes for job titles. The final panel groups by the 11 hiring platform intermediaries observed, with firms that use proprietary platforms included as a single group.

Table 7: Two-way fixed effect estimates of firm components

	Race		Gender		Over 40	
	State	Job title	State	Job title	State	Job title
SD firm effects	0.0176	0.0150	0.0253	0.0255	0.0096	0.0088
SD job title / state effects	0.0003	-	-	0.0080	0.0004	-
Covariance	0.0000	0.0001	0.0000	0.0002	0.0000	0.0002
N jobs	11026	11026	10720	10720	10652	10652
N firms	108	108	108	108	108	108
N job titles / states	51	47	51	47	51	47
N job titles / states > 1 firm	51	43	51	43	51	43
Mean gap	0.0196	0.0196	0.0023	0.0023	0.0037	0.0037
p-value firm effects	0.000	0.0008	0.000	0.000	0.071	.040
p-value job title / state effects	0.186	0.327	0.482	0.237	0.86	0.459

Notes: This table presents bias-corrected variance component estimates from two-way fixed effect models estimated using the leave-out procedure of Kline, Saggio and Sølvsten (2020). Columns labeled "Job title" include fixed effects for the first three digits of each job's O*Net SOC code. Columns labeled "State" include fixed effects for the job's state. All variance and covariance estimates are job-weighted. Only jobs in the leave-job-out connected set are included for each estimate. Dashes indicate negative variance estimates and hence undefined estimated standard deviations. "N job titles / states > 1 firm" is the number of states or job titles in the connected set observed at 2 or more firms. The final two rows report p-values from tests of the joint hypothesis that all firm or job title / state fixed effects equal zero, computed using the heteroscedasticity robust procedure of Anatolyev and Sølvsten (2020).

Table 8: Sensitivity of q-values to estimation strategy

	Race		Gender	Age
	One-tailed	Two-tailed	Two-tailed	Two-tailed
	Bootstrapped λ			
$\hat{\pi}_0$	0.391	0.541	0.833	0.833
# q-values <= 0.05	23	8	1	0
# q-values <= 0.1	45	21	5	1
λ	0.550	0.350	0.300	0.400
	Randomization inference p-values			
$\hat{\pi}_0$	0.324	0.463	0.818	0.787
# q-values <= 0.05	34	24	8	1
# q-values <= 0.1	59	38	10	1
λ	0.600	0.400	0.400	0.400
	Smoothed			
$\hat{\pi}_0$	0.451	0.882	0.854	0.832
# q-values <= 0.05	21	4	1	0
# q-values <= 0.1	40	18	5	1
	95% upper CI for π_0			
$\hat{\pi}_0$	0.607	0.699	1.000	1.000
# q-values <= 0.05	20	4	1	0
# q-values <= 0.1	31	18	5	1

Notes: This table reports the results of estimating firm q-values for discrimination using several strategies. Each panel reports an estimated upper bound on the share of non-discriminating firms (π_0) along with numbers of firms with q-values less than 0.1 and 0.05. Estimates are based on p-values taken from a t-test of mean job-level contact rate gaps for each firm, except in the second panel, which uses p-values constructed based on 10,000 simulations permuting race, gender, and age labels. In accordance with how characteristics were stratified in the experiment, race labels are permuted within pairs, while gender and age are permuted unconditionally. The first two panels estimate π_0 by choosing the tuning parameter λ based on the bootstrap methodology from Storey et al. (2015). The third panel uses the smoothed estimator from Storey (2003). The final panel reports the upper limit of the 95% upper confidence interval for π_0 constructed using the method of Armstrong (2015).

Table 9: Estimates of racial discrimination for firms with q-values below 0.05

		Federal					Posterior	Posterior	Posterior
q-value rank	Industry	Contractor?	Contact gap	Std. err.	p-value	q-value	mean	5th pctile	95th pctile
1	Auto dealers / services	Yes	0.0952	0.0197	0.0000	0.0001	0.0833	0.0439	0.1034
2	Auto dealers / services	No	0.0507	0.0143	0.0003	0.0061	0.0348	0.0133	0.0670
3	Auto dealers / services	No	0.0738	0.0220	0.0005	0.0073	0.0481	0.0190	0.0974
4	Auto dealers / services	No	0.0787	0.0249	0.0010	0.0103	0.0489	0.0199	0.1021
5	Apparel stores	No	0.0733	0.0250	0.0022	0.0158	0.0440	0.0185	0.0917
6	Other retail	No	0.0469	0.0159	0.0020	0.0158	0.0282	0.0118	0.0587
7	Other retail	Yes	0.0605	0.0219	0.0033	0.0176	0.0359	0.0153	0.0731
8	General merchandise	Yes	0.0520	0.0187	0.0031	0.0176	0.0309	0.0131	0.0631
9	Auto dealers / services	No	0.0613	0.0240	0.0060	0.0194	0.0366	0.0157	0.0712
10	Eating/drinking	No	0.0560	0.0222	0.0064	0.0194	0.0334	0.0143	0.0648
11	Other retail	No	0.0560	0.0214	0.0050	0.0194	0.0333	0.0142	0.0658
12	Auto dealers / services	No	0.0540	0.0215	0.0068	0.0194	0.0323	0.0138	0.0623
13	Food stores	Yes	0.0511	0.0204	0.0069	0.0194	0.0305	0.0131	0.0589
14	General merchandise	No	0.0427	0.0170	0.0068	0.0194	0.0255	0.0109	0.0493
15	Furnishing stores	Yes	0.0400	0.0159	0.0066	0.0194	0.0239	0.0102	0.0462
16	Wholesale nondurable	No	0.0386	0.0158	0.0080	0.0199	0.0232	0.0099	0.0442
17	Apparel manufacturing	Yes	0.0350	0.0142	0.0078	0.0199	0.0210	0.0090	0.0401
18	Building materials	Yes	0.0373	0.0157	0.0093	0.0218	0.0226	0.0096	0.0425
19	Health services	Yes	0.0544	0.0240	0.0132	0.0292	0.0335	0.0142	0.0615
20	Furnishing stores	No	0.0400	0.0183	0.0152	0.0322	0.0250	0.0105	0.0452
21	Eating/drinking	No	0.0340	0.0159	0.0172	0.0346	0.0214	0.0090	0.0385
22	General merchandise	No	0.0423	0.0210	0.0229	0.0439	0.0275	0.0114	0.0486
23	Insurance / real estate	No	0.0278	0.0140	0.0257	0.0472	0.0182	0.0075	0.0320

Notes: This table reports estimates of white-Black contact gaps for the 23 individual firms with q-values less than 0.05. P-values and q-values come from one-sided tests of the null hypothesis that the firm does not discriminate against Black applicants. To ensure that q-values are non-decreasing for nested decision thresholds, we follow Storey (2002, 2003) in estimating $\hat{q}_f$ as $\min_{t \geq \hat{p}_f} \widehat{FDR}(t)$, which implies firms with different p-values may have the same q-value. Posterior means and percentiles are empirical Bayes posteriors constructed using the estimated distribution in Figure 7 as the prior.

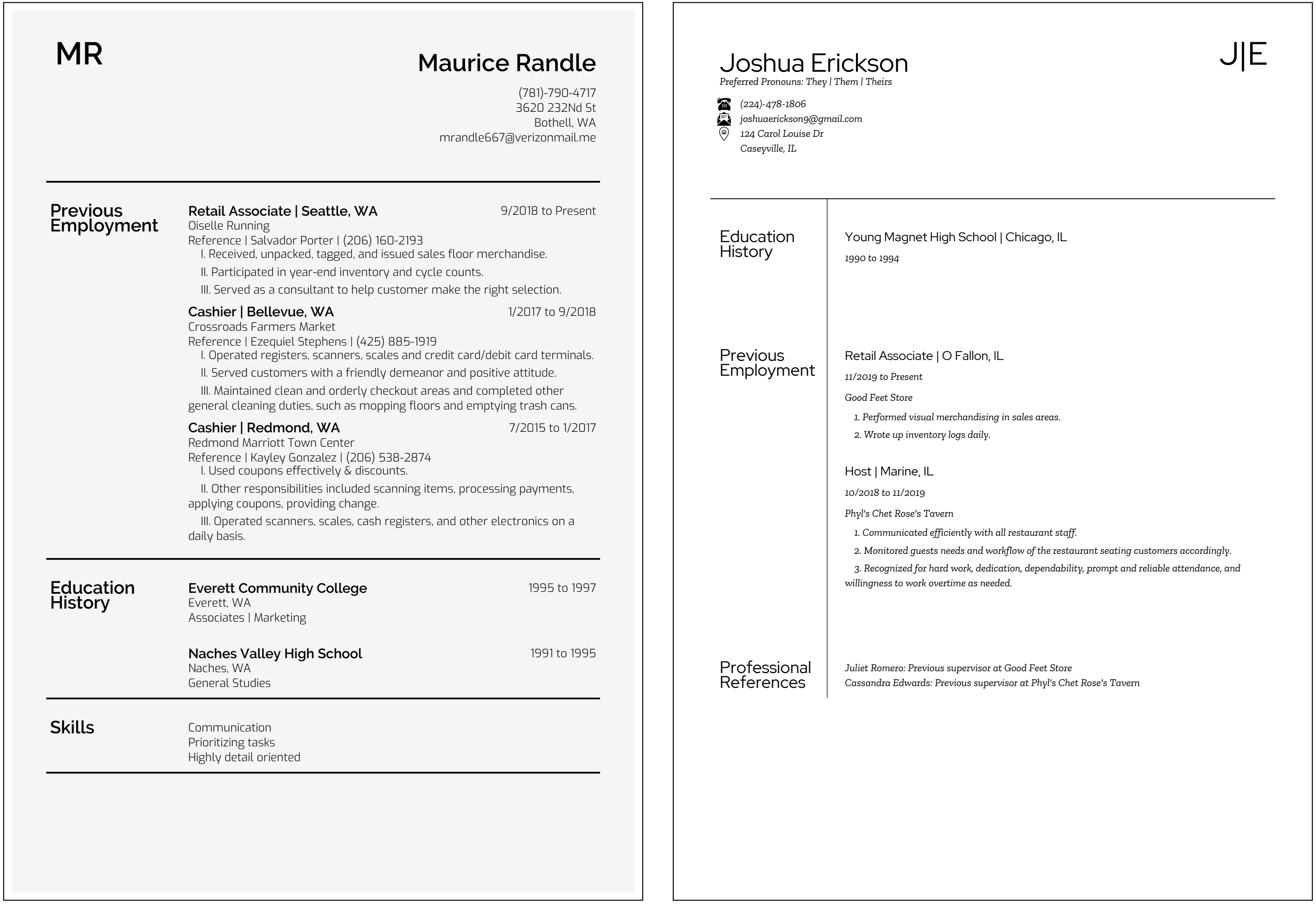
Table 10: Discrimination estimates and detection by industry

			Race			Gender		
SIC	Industry	N firms	W-B post gap	# q-val < .05	Mean LFDR	M-F post gap	# q-val < .05	Mean LFDR
20	Food products	2	0.015	0	1.000	-0.004	0	1.000
23	Apparel manufacturing	2	0.020	1	0.197	0.007	0	0.719
24-35	Other manufacturing	4	0.018	0	0.417	0.011	0	0.679
42-47	Freight / transport	4	0.011	0	0.910	0.001	0	0.952
48	Communications	2	0.017	0	0.393	0.013	0	0.984
49	Electric / gas	3	0.015	0	0.392	0.002	0	0.988
50	Wholesale durable	2	0.017	0	0.339	0.034	0	0.556
51	Wholesale nondurable	11	0.018	1	0.512	0.005	0	0.875
52	Building materials	3	0.014	1	0.614	0.012	0	0.853
53	General merchandise	12	0.022	2	0.319	-0.001	0	0.878
54	Food stores	5	0.025	1	0.411	0.009	0	0.826
55	Auto dealers / services	8	0.039	6	0.147	0.005	0	0.887
56	Apparel stores	4	0.025	1	0.292	-0.060	1	0.420
57	Furnishing stores	4	0.022	2	0.351	-0.006	0	0.800
58	Eating/drinking	5	0.027	2	0.350	0.003	0	0.934
59	Other retail	7	0.021	3	0.363	-0.002	0	0.978
60-61	Banks / credit	5	0.010	0	0.720	0.002	0	0.798
62	Securities brokers	2	0.010	0	0.473	-0.011	0	0.657
63-65	Insurance / real estate	8	0.013	0	0.521	-0.003	0	0.922
70	Accommodation	2	0.015	0	0.569	0.001	0	1.000
73	Business services	3	0.012	0	0.612	0.000	0	0.949
75-76	Auto / repair services	3	0.013	0	0.521	0.015	0	0.640
80	Health services	5	0.016	1	0.791	-0.009	0	0.922
87	Engineering services	2	0.009	0	0.401	-0.001	0	0.977

Notes: This table shows the results of aggregating firm-specific posterior estimates of race and gender discrimination to the industry level. Industries that include only one firm are grouped together with proximate SIC codes. The column "W-B post gap" shows industry averages of posterior mean white/Black contact gaps. The column "M-F post gap" displays industry averages of posterior mean male/female contact gaps. The column "# q-val < .05" gives the number of firms in the industry with q-values below 0.05. The column "mean LFDR" reports the mean Local False Discovery Rate (LFDR) among firms in the industry. Firm level q-values and LFDRs were estimated using the procedure of Storey et al. (2015). The distribution of race LFDRs is depicted in Panel (a) of Figure 11. The distribution of gender LFDRs is depicted in Panel (c) of Figure 11.

Appendix A Additional Figures and Tables

Figure A1: Examples of applicant resumes



MR

Maurice Randle

(781)-790-4717
3620 232Nd St
Bothell, WA
mrandle667@verizonmail.me

Previous Employment

Retail Associate | Seattle, WA 9/2018 to Present
Oiselle Running
Reference | Salvador Porter | (206) 160-2193
I. Received, unpacked, tagged, and issued sales floor merchandise.
II. Participated in year-end inventory and cycle counts.
III. Served as a consultant to help customer make the right selection.

Cashier | Bellevue, WA 1/2017 to 9/2018
Crossroads Farmers Market
Reference | Ezequiel Stephens | (425) 885-1919
I. Operated registers, scanners, scales and credit card/debit card terminals.
II. Served customers with a friendly demeanor and positive attitude.
III. Maintained clean and orderly checkout areas and completed other general cleaning duties, such as mopping floors and emptying trash cans.

Cashier | Redmond, WA 7/2015 to 1/2017
Redmond Marriott Town Center
Reference | Kayley Gonzalez | (206) 538-2874
I. Used coupons effectively & discounts.
II. Other responsibilities included scanning items, processing payments, applying coupons, providing change.
III. Operated scanners, scales, cash registers, and other electronics on a daily basis.

Education History

Everett Community College 1995 to 1997
Everett, WA
Associates | Marketing

Naches Valley High School 1991 to 1995
Naches, WA
General Studies

Skills

Communication
Prioritizing tasks
Highly detail oriented

Joshua Erickson

J|E

Preferred Pronouns: They | Them | Theirs

☎ *(224)-478-1806*
✉ *joshuaerickson9@gmail.com*
📍 *124 Carol Louise Dr*
Caseyville, IL

Education History

Young Magnet High School | Chicago, IL
1990 to 1994

Previous Employment

Retail Associate | O Fallon, IL
11/2019 to Present
Good Feet Store
1. Performed visual merchandising in sales areas.
2. Wrote up inventory logs daily.

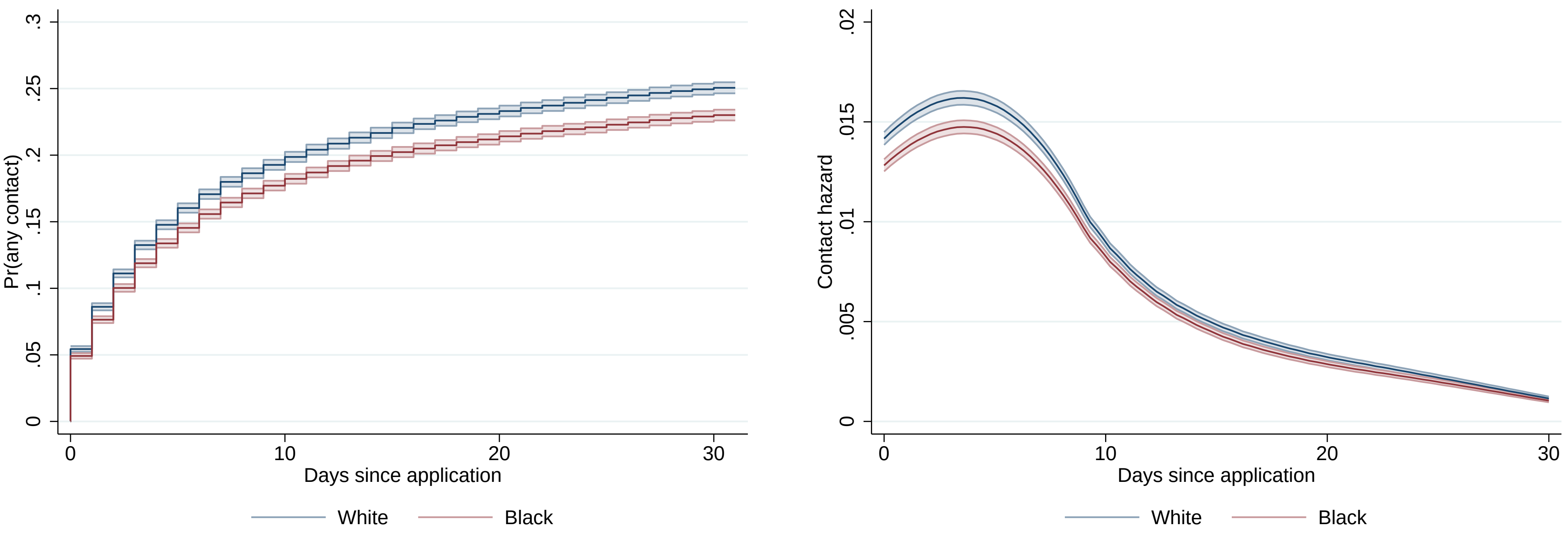
Host | Marine, IL
10/2018 to 11/2019
Phyl's Chet Rose's Tavern
1. Communicated efficiently with all restaurant staff.
2. Monitored guests needs and workflow of the restaurant seating customers accordingly.
3. Recognized for hard work, dedication, dependability, prompt and reliable attendance, and willingness to work overtime as needed.

Professional References

Juliet Romero: Previous supervisor at Good Feet Store
Cassandra Edwards: Previous supervisor at Phyl's Chet Rose's Tavern

Notes: This figure presents two examples of randomly generated resumes used in the experiment. Resumes are formatted using a combination of pre-set options specifying length, fonts, text sizes, section header names, and layouts, with controls to ensure resumes that overflow one page are not generated. The resume on the right features gender-neutral pronouns displayed below the name.

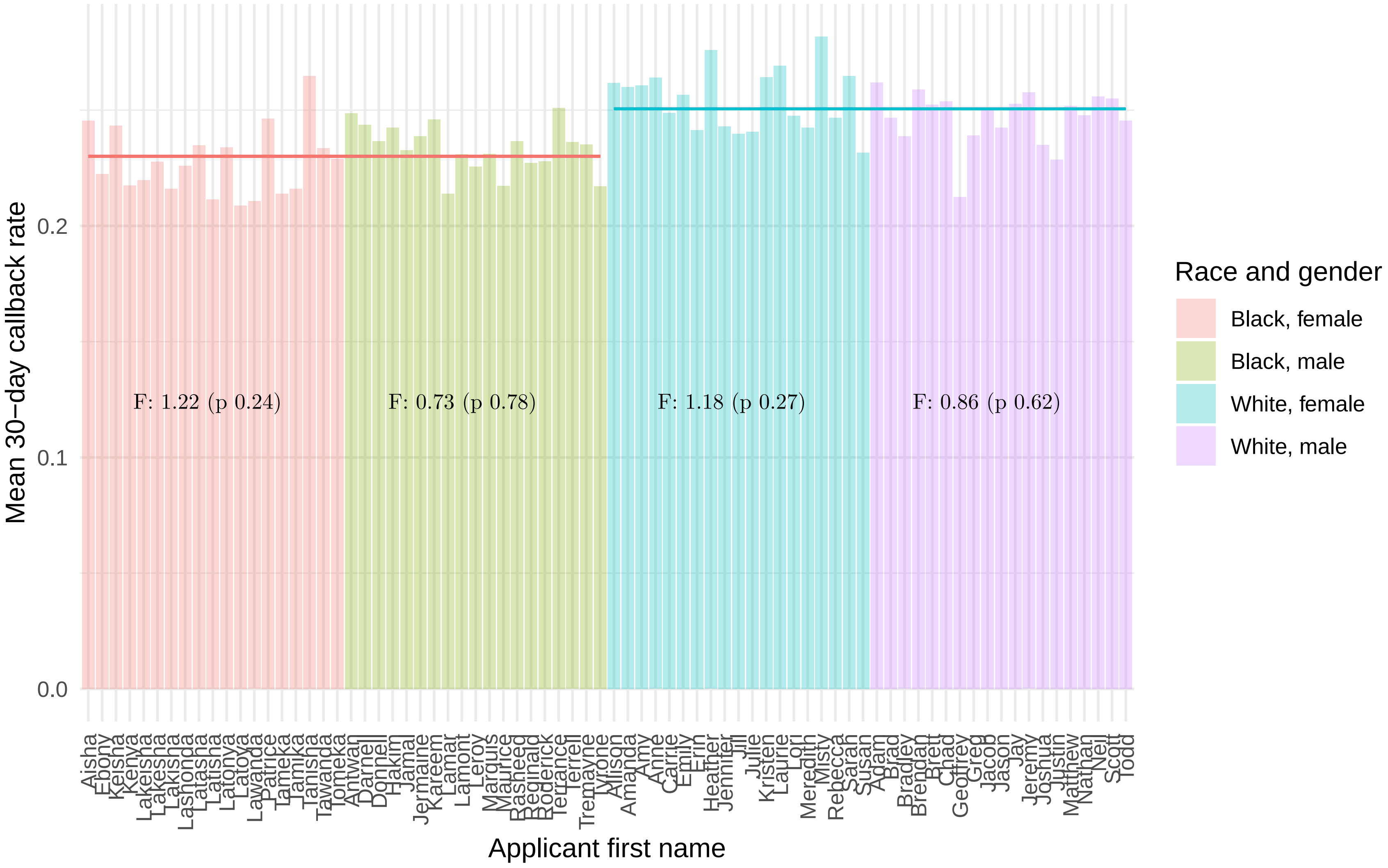
Figure A2: Kaplan-Meier estimates of contact probability and smoothed hazard


a) Contact probability

b) Smoothed contact hazard

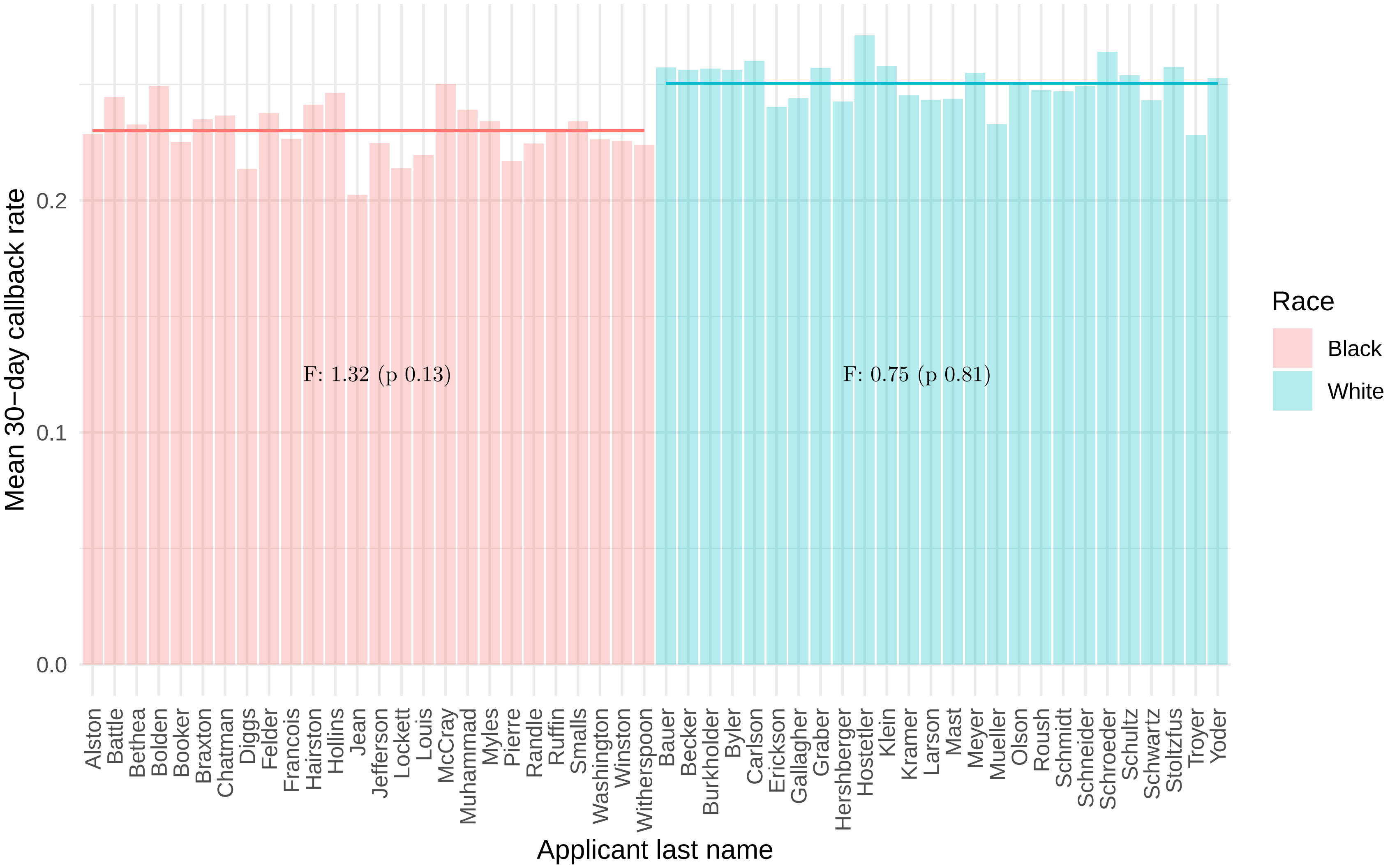
Notes: This figure plots contact probabilities and hazards as functions of days since application. Contact probabilities correspond to Kaplan-Meier failure function estimates. Hazards are Kaplan-Meier hazard estimate smoothed using the Epanechnikov kernel. Shaded areas represent pointwise 95% confidence bands.

Figure A3: Callbacks by applicant first name



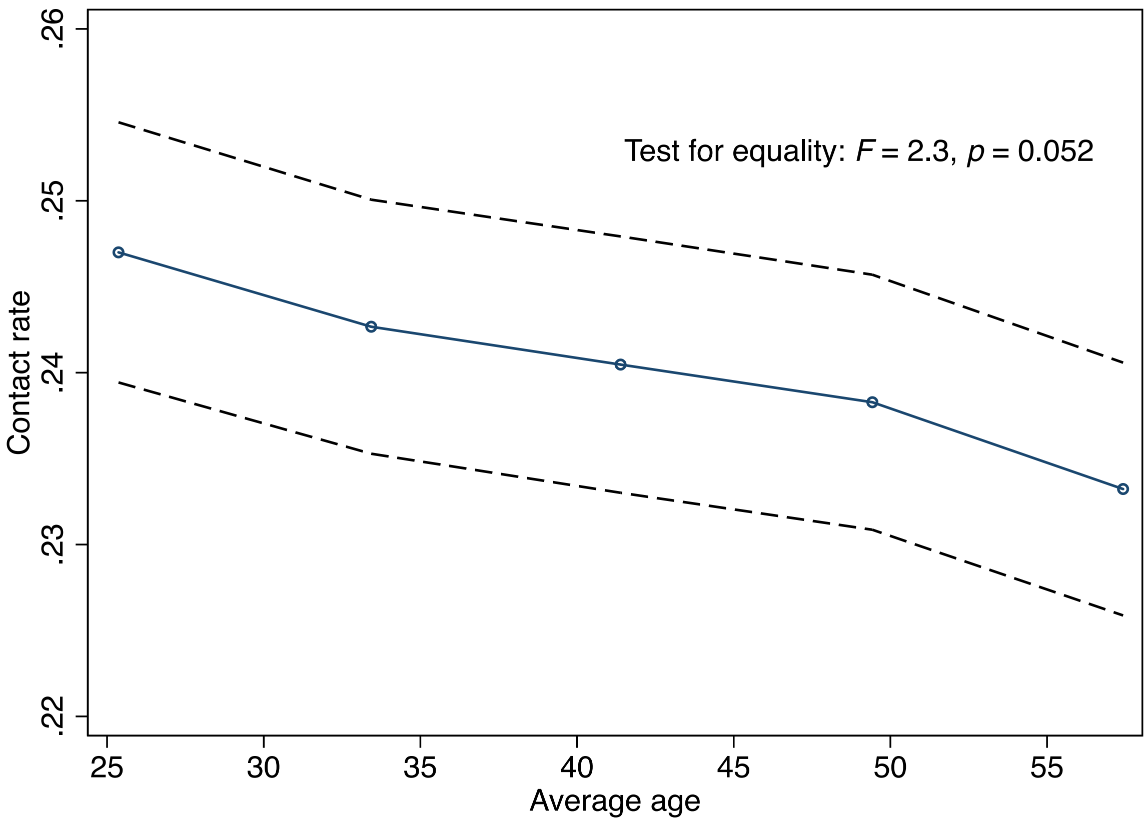
Notes: This figure shows mean contact rates by applicant first name, organized by race and gender group. The horizontal bars show race group mean contact rates. F-tests and p-values come from joint tests of the hypothesis that contact rates are equal across names separately by race and gender group.

Figure A4: Callbacks by applicant last name



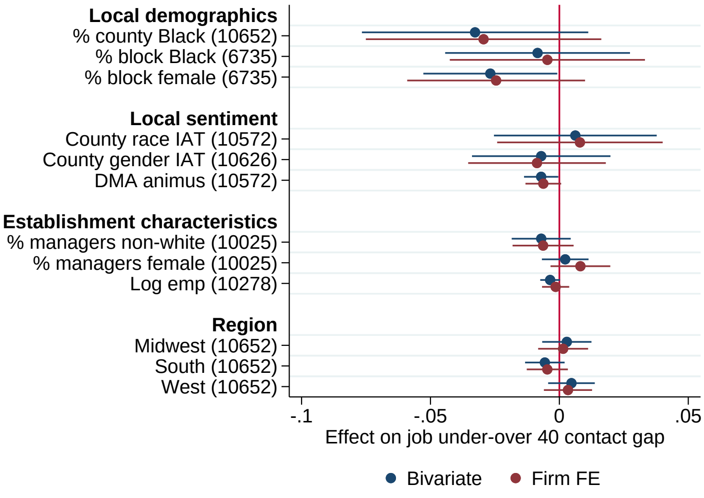
Notes: This figure shows mean contact rates by applicant last name, organized by race. The horizontal bars show race group mean contact rates. F-tests and p-values come from joint tests of the hypothesis that contact rates are equal across names separately by race.

Figure A5: Contact rates by age category



Notes: This figure plots average 30-day contact rates by quintile of applicant age at the time of application. Estimates come from regressions of a contact indicator on indicators for age quintile, controlling for wave indicators. The horizontal axis plots average age in each quintile. The vertical axis plots the mean contact rate, calculated as the sum of the quintile coefficient and mean wave effect. Dashed lines indicate 90% confidence intervals. F-statistic and p-value come from a Wald test that contact rates are equal across quintiles, clustering standard errors by job.

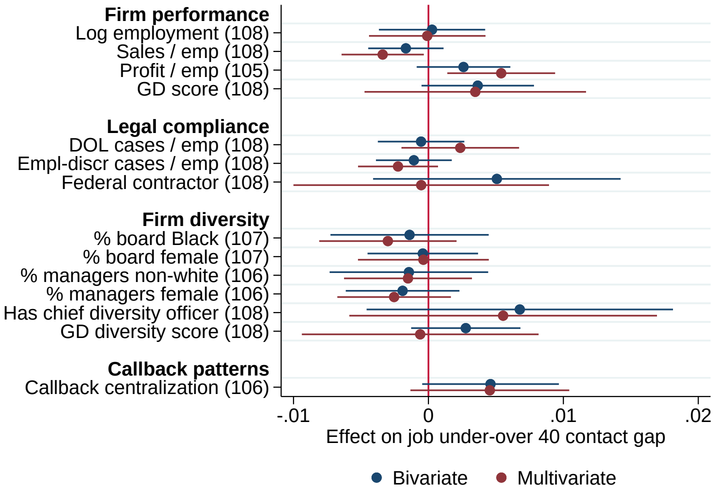
Figure A6: Relationships between age contact gaps and establishment characteristics



P-value for joint sig w/o firm FE: 0.64, w/ firm fe: 0.67

Notes: This figure plots the relationship between establishment-level covariates and contact gaps for applicant age under vs. over 40. Each relationship is estimated with a linear regression with job-level contact gaps as the outcome. All jobs with defined contact gaps for age and matched to the listed covariate are included. "Bivariate" points plot coefficients from a regression of contact gaps on the covariate alone. "Firm FE" points include firm fixed effects. Bars indicate 95% confidence intervals based on robust standard errors. Appendix C provides a complete description of covariate definitions and sources.

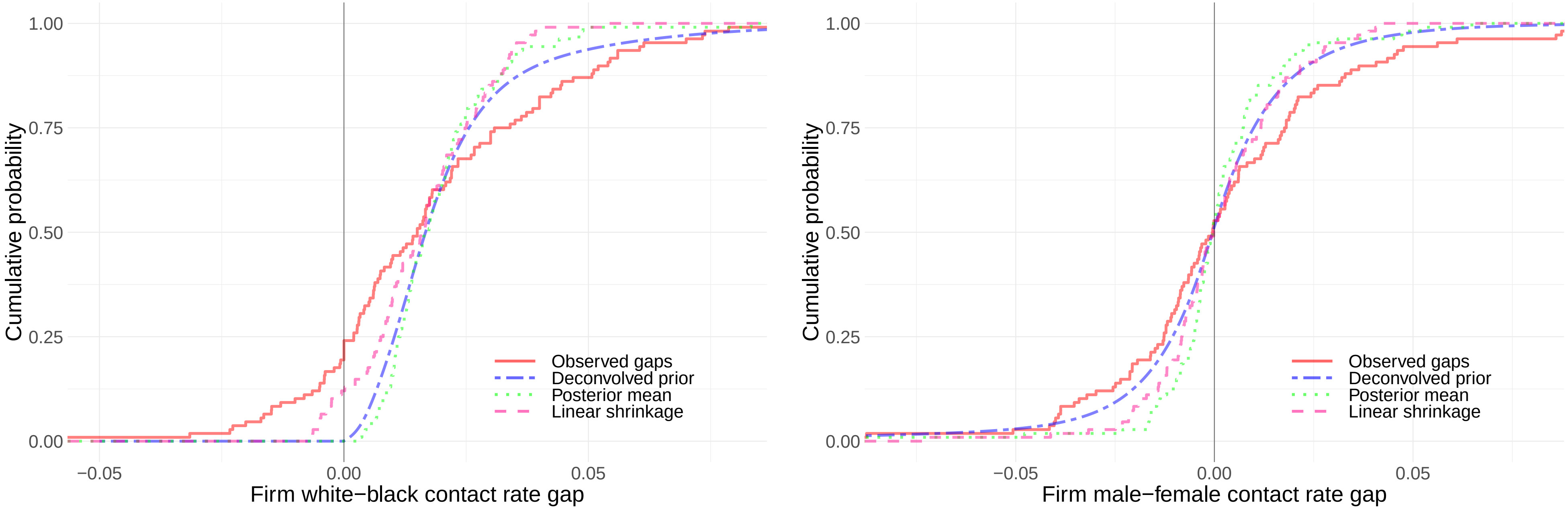
Figure A7: Relationships between age contact gaps and firm characteristics



P-value for joint significance: 0.341

Notes: This figure plots relationships between firm-level covariates and contact gaps for application age under vs. over 40. Each relationship is estimated with a linear regression with job-level contact gaps as the outcome. All covariates are standardized to be mean zero, standard deviation 1 in sample."Bivariate" points plot coefficients from a regression of contact gaps on the covariate alone. "Multivariate" points plot effects when all covariates are entered simultaneously. Bars indicate 95% confidence intervals based on standard errors clustered at the firm level. Appendix C provides a complete description of covariate definitions and sources.

Figure A8: Distribution of observed, deconvolved, and posterior estimates

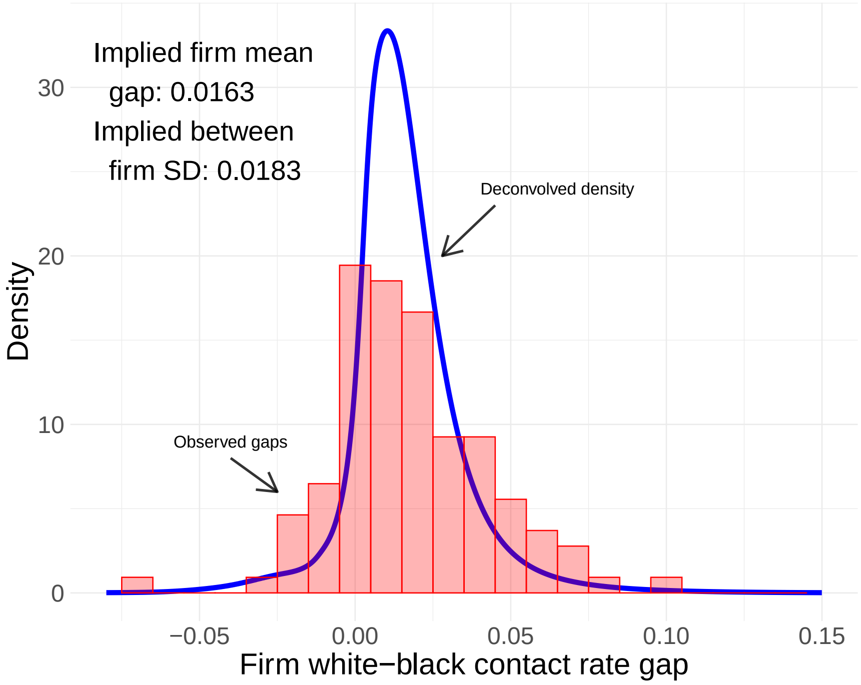


a) Race

b) Gender

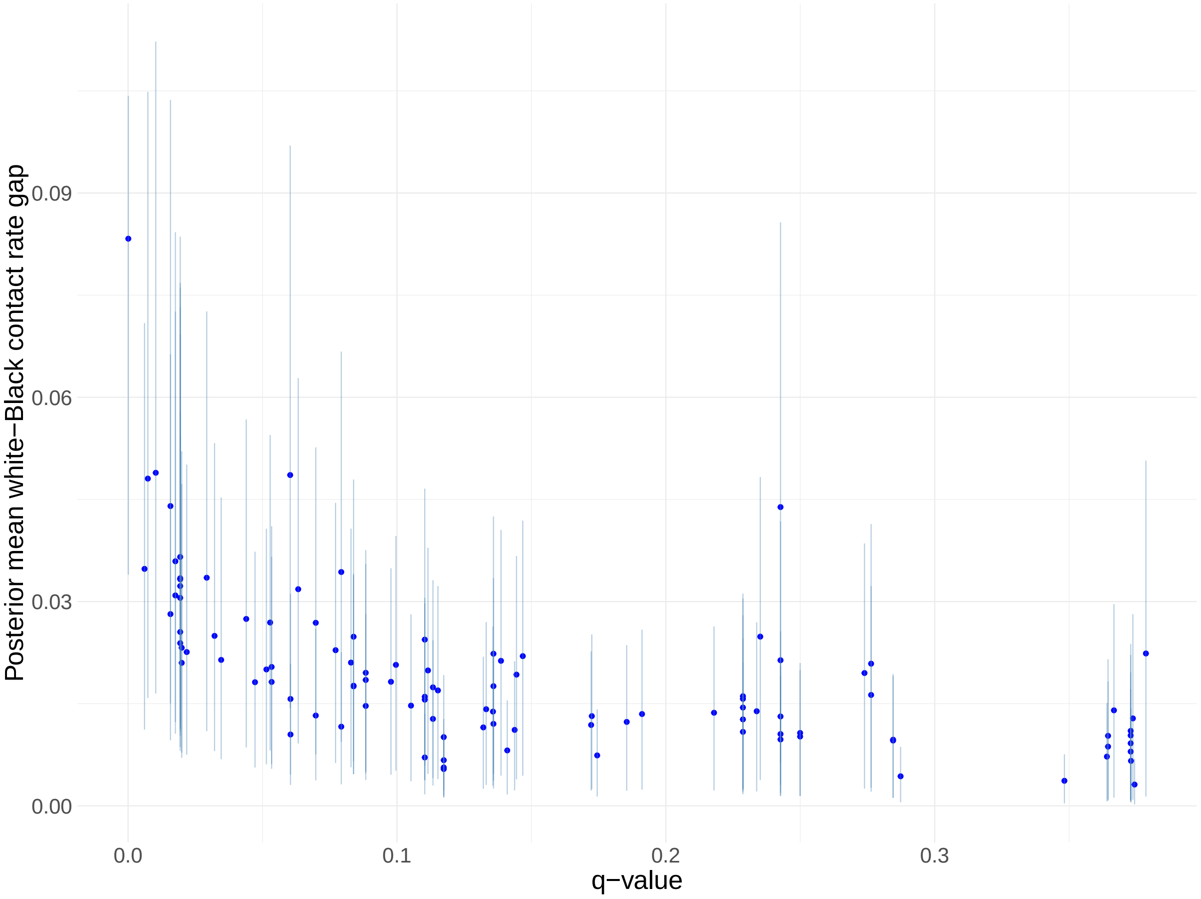
Notes: This figure presents estimates of the distribution of firm-specific contact gaps for race and gender. The red solid line presents the empirical CDF of estimated gaps. The blue dashed line plots the CDF of population contact gaps based on the deconvolution estimates in Figure 7. The green dotted line plots the empirical CDF of posterior means, constructed treating the deconvolved density as prior knowledge. The pink dashed line shows the empirical CDF of estimates shrunk linearly towards the grand mean with weights given by the signal-to-noise ratio $\hat{\theta}/(s_f^2 + \hat{\theta})$, where $\hat{\theta}$ is the bias-corrected estimate of the variance of contact gaps across firms.

Figure A9: Deconvolution of firm-level racial discrimination without support restriction



Notes: This figure presents non-parametric estimates of the distribution of firm-specific white-Black contact rate differences. The red histogram shows the distribution of estimated firm contact gaps. Blue line shows estimates of the population contact gap distributions. The population distributions are estimated by applying the deconvolveR package (Narasimhan and Efron, 2020) to firm-specific z-score estimates, then numerically integrating over the distribution of log standard errors using a Gaussian kernel and the Silverman (1986) bandwidth to recover the distribution of contact gaps. The penalization parameter in the deconvolution step is calibrated so that the resulting distribution matches the corresponding bias-corrected variance estimate from Table 4.

Figure A10: Posterior mean contact gaps vs. q-values



Notes: This figure plots posterior mean white/Black contact gaps $\bar{\Delta}_f$ for each firm against estimated q-values for racial discrimination. Vertical lines depict 95% empirical Bayes credible intervals (EBCIs).

Table A1: Balanced sample: Firm-level heterogeneity in discrimination

	Contact gap SD				
	(1) Bias-corrected	(2) Cross-Wave	(3) Cross-State	(4) χ^2 test of heterogeneity	(5) p-value for no discrim against:
Race	0.0184	0.0171	0.0182	229.5	W: 1.00
	(0.0030)	(0.0033)	(0.0031)	[0.000]	B: 0.00
Gender	0.0207	0.0213	0.0200	124.2	M: 0.06
	(0.0041)	(0.0043)	(0.0043)	[0.000]	F: 0.03
Over 40	0.0098	0.0096	0.0099	90.2	Y: 0.14
	(0.0045)	(0.0050)	(0.0045)	[0.072]	O: 0.02

Notes: This table presents estimated standard deviations of firm-level contact rate gaps and tests for heterogeneity in gaps using the balanced sample of firms present in all five waves. Column 1 reports bias-corrected estimates of standard deviations of average contact gaps across firms based on covariances between job-level contact gaps. Columns 2 and 3 report cross-wave and cross-state estimates based on covariances between firm-by-wave and firm-by-state contact gaps. Details on these estimators appear in the Appendix. Standard errors for all variance estimators are produced by job-clustered weighted bootstrap. Column 4 displays χ^2 values and associated p-values from tests of the null hypothesis of no heterogeneity in discrimination. The test statistic is $\sum_f (\hat{\Delta}_f - \bar{\Delta})^2/s_f^2$, where $\hat{\Delta}_f$ is the contact cap estimate for firm f, s_f is the estimate's standard error, and $\bar{\Delta}$ is the equally-weighted average of contact gaps. Column 5 presents one-sided tests of no discrimination against white (W), black (B), male (M), female (F), aged under 40 (Y), and over 40 (O) applications using the methodology in Bai, Santos and Shaikh (2021).

Table A2: Variance components for other resume attributes

	Contact gap SD				
	(1) Bias-corrected	(2) Cross-Wave	(3) Cross-State	(4) χ^2 test of heterogeneity	(5) p-value for no discrim against:
LGBTQ Club Member	-	-	-	88.0	Y: 1.00
				[0.885]	N: 0.98
Gender-Neutral Pronouns	0.0198	0.0177	0.0147	126.5	Y: 0.92
	(0.0111)	(0.0126)	(0.0138)	[0.076]	N: 0.65

Notes: This table presents estimated standard deviations of firm-level contact rate gaps by LGBTQ club member status and the presence of gender-neutral pronouns, along with tests for heterogeneity in these gaps. Column 1 reports bias-corrected estimates of standard deviations of average contact gaps across firms based on covariances between job-level contact gaps. Columns 2 and 3 report cross-wave and cross-state estimates based on covariances between firm-by-wave and firm-by-state contact gaps. Details on these estimators appear in the Appendix. Standard errors for all variance estimators are produced by job-clustered weighted bootstrap. Column 4 displays χ^2 values and associated p-values from tests of the null hypothesis of no heterogeneity in discrimination. The test statistic is $\sum_f (\hat{\Delta}_f - \bar{\Delta})^2/s_f^2$, where $\hat{\Delta}_f$ is the contact cap estimate for firm f, s_f is the estimate's standard error, and $\bar{\Delta}$ is the equally-weighted average of contact gaps. Column 5 presents one-sided tests of no discrimination against applicants with the relevant attribute (Y) and those without the attribute (N) using the methodology in Bai, Santos and Shaikh (2021). Estimates include all 108 firms.

Table A3: Relationship between z-scores and standard errors

	Race		Gender	
	(1) Full sample	(2) Split sample	(3) Full sample	(4) Split sample
Z-score	33.98	18.06	11.50	4.52
	(24.07)	(11.35)	(14.12)	(6.74)
Squared residual	86.20	17.94	83.17	28.78
	(48.44)	(17.58)	(53.30)	(16.94)

Notes: This table assesses dependence between firm-specific z-score estimates and standard errors. Coefficients in the first row come from regressions of z-scores on standard errors, and coefficients in the second row come from regressions of the squared residuals from the first row on standard errors. Columns 1 and 2 display results for race, and columns 3 and 4 show results for gender. Columns 1 and 3 use z-scores and standard errors computed in the full sample of jobs. Columns 2 and 4 randomly split the jobs at each firm into two equally-sized samples and regress z-scores computed in one sample on standard errors computed in the other sample, stacking the two samples and clustering standard errors by firm identifier.

Table A4: Job-level discrimination prevalence bounds

	(1) Race	(2) Gender	(3) Over 40
Mean gap	0.020	0.002	0.004
	(0.002)	(0.002)	(0.002)
Total job-level variance	0.071	0.093	0.018
	(0.000)	(0.000)	(0.000)
Prevalence bound	0.070	0.000	0.027
	(0.011)	(0.001)	(0.037)

Notes: This table reports a bound on the job-level prevalence of discrimination assuming that a fixed fraction of jobs discriminate and the remaining jobs exhibit contact gaps of zero. The mean gap reported is the job-weighted average contact gap. The total job level variance is computed as the covariance of contact gaps among the first four and last four applications at every job. The prevalence bound is estimated as $(\hat{\Delta}^2 - s^2)/(\hat{\sigma}^2 + \hat{\Delta}^2 - s^2)$, where $\hat{\Delta}^2$ is the square of the estimated mean gap, s is the mean gap's estimated standard error, and $\hat{\sigma}^2$ is the estimated between-job variance. Bootstrap standard errors are reported in parentheses.

Appendix B Details of Experimental Design

Resume characteristics

Names: We draw racially distinctive first names from two sources. First, we use the same set of names in Bertrand and Mullainathan (2004), which are in turn drawn from Massachusetts birth records covering 1974 to 1979. Second, we supplement with names drawn from administrative records on speeding infractions and arrests provided by the North Carolina Administrative Office of the Courts and covering 2006 to 2018. We pick the most common names among drivers born between 1974 and 1979 with race- and gender-specific shares of at least 90%. The top names using this criterion substantially overlaps with Bertrand and Mullainathan (2004)'s list, with 6/9, 4/9, 4/9, and 3/9 names included in both sources for Black women, Black men, white women, and white men, respectively. We add 10 new names from the N.C. records for each race and gender group, leaving 19 total first names per group.

Table B1: First names assigned by race and gender

	Black male		White male		Black female		White female	
	Name	Source	Name	Source	Name	Source	Name	Source
1	Antwan	NC	Adam	NC	Aisha	Both	Allison	BM
2	Darnell	BM	Brad	Both	Ebony	Both	Amanda	NC
3	Donnell	NC	Bradley	NC	Keisha	BM	Amy	NC
4	Hakim	BM	Brendan	Both	Kenya	BM	Anne	BM
5	Jamal	Both	Brett	BM	Lakeisha	NC	Carrie	BM
6	Jermaine	Both	Chad	NC	Lakesha	NC	Emily	Both
7	Kareem	Both	Geoffrey	BM	Lakisha	Both	Erin	NC
8	Lamar	NC	Greg	BM	Lashonda	NC	Heather	NC
9	Lamont	NC	Jacob	NC	Latasha	NC	Jennifer	NC
10	Leroy	BM	Jason	NC	Latisha	NC	Jill	Both
11	Marquis	NC	Jay	BM	Latonya	Both	Julie	NC
12	Maurice	NC	Jeremy	NC	Latoya	Both	Kristen	Both
13	Rasheed	BM	Joshua	NC	Lawanda	NC	Laurie	BM
14	Reginald	NC	Justin	NC	Patrice	NC	Lori	NC
15	Roderick	NC	Matthew	Both	Tameka	NC	Meredith	BM
16	Terrance	NC	Nathan	NC	Tamika	Both	Misty	NC
17	Terrell	NC	Neil	BM	Tanisha	BM	Rebecca	NC
18	Tremayne	BM	Scott	NC	Tawanda	NC	Sarah	Both
19	Tyrone	Both	Todd	BM	Tomeka	NC	Susan	NC

Notes: This table lists the first names assigned by race and gender and their sources. "BM" indicates that the name appeared in original set of nine names used for each group in Bertrand and Mullainathan (2004). "NC" indicates the name was drawn from data on North Carolina speeding infractions and arrests. "Both" indicates the name appeared in both sources. Names from N.C. speeding tickets were selected from the most common names where at least 90% of individuals are reported to belong to the relevant race and gender group.

Last names are drawn from 2010 Decennial Census data. We use the names with highest race-specific shares that occur at least 10,000 times, picking 26 total for each race group. Each resume is assigned a first and last name from the appropriate race and gender group, sampling without replacement within firm. Each pair of applicants was assigned a white and Black first and last name, with the gender of the first name chosen randomly.

Table B2: Last names assigned by race

	Black			White		
	Name	Frequency	Race share	Name	Frequency	Race share
1	Alston	30,693	79.8	Bauer	65,004	95.1
2	Battle	26,432	77.3	Becker	87,859	94.89
3	Bethea	12,061	74.8	Burkholder	11,532	97.55
4	Bolden	21,819	72.3	Byler	13,230	98.19
5	Booker	36,840	65.2	Carlson	120,552	94.83
6	Braxton	12,268	72.4	Erickson	82,085	95.05
7	Chatman	15,473	79.2	Gallagher	69,834	94.62
8	Diggs	14,467	68.1	Graber	12,204	97.16
9	Felder	13,257	66.9	Hershberger	14,357	98.08
10	Francois	14,593	78	Hostetler	14,505	97.46
11	Hairston	16,090	80.9	Klein	81,471	95.41
12	Hollins	10,213	73.8	Kramer	63,936	95.35
13	Jean	21,140	70.3	Larson	122,587	94.79
14	Jefferson	55,179	74.2	Mast	15,932	96.99
15	Lockett	14,140	71.4	Meyer	150,895	94.84
16	Louis	23,738	65.5	Mueller	64,191	95.66
17	McCray	28,024	67.4	Olson	164,035	94.76
18	Muhammad	19,076	82.9	Roush	11,386	96.44
19	Myles	13,898	72.1	Schmidt	147,034	95.15
20	Pierre	33,913	86.7	Schneider	101,290	95.35
21	Randle	14,437	68.8	Schroeder	67,977	95.36
22	Ruffin	16,324	80.4	Schultz	104,888	94.81
23	Smalls	12,435	90.5	Schwartz	90,071	95.93
24	Washington	177,386	87.5	Stoltzfus	15,786	99
25	Winston	21,667	62.7	Troyer	16,981	97.96
26	Witherspoon	13,171	62.1	Yoder	56,410	97.77

Notes: This table reports the last names used in the experiment. Names are drawn from Decennial Census data. We pick names with the highest race-specific shares among those that occur more than 10,000 times. The table reports each name's frequency and the share of individuals with that surname who belong to each race group.

Dates of birth: Applicants were initially randomly assigned a date of birth between 1960 and 2000. Because these dates were fixed, as the experiment continued the average age of applicants increased. In wave 5 we began to assign dates of birth implying a uniform distribution of applicant ages between 20 and 60 at the time of application creation.

Social security numbers: Some applications required us to provide a social security number. We assigned all applicants a social security number from a publicly available database of numbers belonging to the deceased.

Emails: We manually created Gmail, Outlook, and Yahoo email accounts for roughly half of our applicants. To facilitate account creation and avoid account limits on these service, we also registered new domains designed to imitate common internet service providers' names: icloudlive.me, spectrumemail.org, fiosmail.net, and xfinity19.com. Each domain redirected to the relevant provider (e.g., icloudlive.me redirected to the icloud home page). Email addresses were creating using combinations of assigned first and last names and random integers. Each email was associated with a single first and last name combination. All emails were set up to automatically forward to a single inbox that was monitored for contacts.

Phone numbers: We provisioned phone numbers from Twilio. During each wave of the experiment, we rented roughly 200 numbers with SMS capabilities from area codes across the country. Each number was assigned to a single first and last name combination, ensuring that the same number was used only once at each company. We rented new numbers each wave so that each unique number was used at each firm at most once.

Phone calls to each number were automatically directed to a voicemail with a standard, non-personalized message. All calls were logged. Any voicemails were recorded and transcribed. We then used a combination of manual and automatic methods to tag voicemails as callbacks from particular employers using text searches on transcribed voicemails and by listening to voicemails. Text message callbacks were processed in the same way.

Addresses: We assigned each application a home address close to the job to which the application was submitted. Addresses were sourced from openaddresses.io and the U.S. Department of Transportation's National Address Database. We download the full set of addresses from both sources and manually eliminated unusual and non-residential addresses. Addresses were randomly assigned to applications without replacement for each job from the set of addresses in zip codes within 20 miles of the target job. If insufficient addresses were available with a 20 mile radius, a 40 mile radius was used instead.

Educational history: All applicants were assigned a high school in same state as the target job. We use the National Center for Educational Statistics to identify all non-specialized public schools with instruction in grades 9-12 and randomly select a school from zip codes with an absolute difference of less than 1,000 from the target job's zip code. If insufficient schools are available, we randomly assign a school from anywhere in the state. All applicants graduated from high school the same year they turned 18 years old.

We attempted to randomly assign half of our applicants an associate degree from a community college in the same state as the target job. We use the Department of

Education's College Scorecard data to identify all relevant degree-granting institutions, manually eliminating some specialty schools. Colleges were assigned in the same manner as high schools. Each applicant with a degree was also assigned a major from a list of common, non-specialized degrees, including Business Technology, Marketing, Information Technology, Communication Studies, and Sales Management. All applicants received their degree two years after finishing high school. Because appropriate colleges were not available in all geographies, slightly less than half of applicants were assigned a degree.

Club membership: Beginning in Wave 2, 20% of applicants were assigned a club to be listed on their resume as part of their educational experience. Half of applicants assigned a club listed clubs intended to signal LGBTQ affiliation: the Gay-Straight Alliance, the Lesbian, Gay, Bisexual, Transgender, and Queer Association, and the Queer-Straight Alliance. The remaining half were assigned either a generic club (History Club, Speech and Debate Club, Foreign Language Club, Outdoors Club, Model United Nations, Performing Arts Club, Student Government, or Music Club) or political club (Young Republican Club, Student Republican Association, Young Republican Club, Student Republican Association, Young Democrat Club, Student Democrat Association, Young Democrat Club, or Student Democrat Association). Applicants were randomly listed as the president, founder, secretary, vice-president or member of the assigned club.

Pronouns: Beginning in Wave 2, 10% of applicants were assigned preferred pronouns. Half of applicants with pronouns received gender-neutral pronouns (they/them/their), and half received pronouns reflecting the typical gender identity of their first name (he/him/his or she/her/hers). Pronouns were listed on the PDF resumes near name and contact information.

Employment history: Each applicant was assigned two to three previous employers. Employers were drawn from the universe of establishment names and addresses listed in the Reference USA dataset. As with addresses, we sample previous employers from zip codes within 20 miles of the target job's zip code, or 40 miles if insufficient employers are available within 20. We exclude any establishments from the same firm as the target job.

Each target job was assigned one of four employment categories: general, retail, clerical, and manual labor. Applicants to general category jobs were assigned previous employers from SIC codes 15, 24, 25, 34, 36, 42, 53, 54, 56, 58, 64, 65, 70, 73, and 80. Applicants to retail category jobs were assigned previous employers from codes 53, 54, 56, 58 and 70. Applicants to clerical jobs were assigned previous employers from codes 15, 24, 25, 34, 36, 64, 65, 73, and 80. Applicants to manual labor jobs were assigned previous employers from codes 34, 36, 25, 24, 15, and 42. Prior employers were assigned without replacement for all applications to the same target job.

Entry-level job titles were assigned for each previous employer appropriate to the industry and experience. Jobs at retail establishments were assigned job titles from Team Member, Retail Associate, Cashier, Stocker, and Customer Service Associate. Jobs at fast-food / quick-service restaurants were assigned titles from Crew Member, Cashier, Food prep / service, and Cook. Jobs at restaurants were assigned titles from Server, Dishwasher, Cashier, Host, and Cook. Jobs at manufacturers and wholesalers were assigned titles from Package Handler, Handler, Laborer, Delivery Driver / Courier, Dockworker, and Warehouse Associate. Office and clerical positions were assigned titles from Office Manager, Receptionist, and Assistant. Jobs at hotels were assigned titles of Housekeeper or Receptionist.

Each job was assigned a fictional supervisor with a first and last name drawn from the most common in the United States and a fictional phone number. Since some applications required us to list a reason for leaving each previous job, we populated a large list of sample reasons (e.g., insufficient hours, seeking promotion opportunity, etc.) and randomly assigned them to each previous job.

Tenure in previous jobs was selected uniformly from 9 to 24 months. No interruptions in employment history were assigned and all applicants reported being currently employed by their most recent prior employer.

We assigned a sample of two to three job duties scraped from online databases of resumes such as jobhero.com. We manually cleaned and formatted these duties to eliminate references to specific employer names or technologies. Duties were entered into "responsibilities / duties" sections of target job applications.

References: When required, applicants listed references using the fictional supervisors at their previous employers.

Personality and skills assessments: Some jobs required applicants to complete personality or skills assessments before they could be considered for an interview. We developed guides for each of these assessments that randomly specified acceptable answers within a range appropriate for the question. Our answers avoided providing an obviously negative signal about applicant quality (e.g., answering "Yes" to "Is it ever acceptable to steal from an employer?"). When questions had no obvious connection to applicant quality, we answered randomly but ensured that answers remained consistent across questions. We answered analytical-reasoning and skill-based questions to mimic the performance of our undergraduate volunteers.

Miscellaneous resume characteristics: Many applications required answering a large number of idiosyncratic questions, ranging from open-ended questions about why the applicant wants to work at the target employer to questions about willingness to comply with employer rules about dress, drug use, and conduct. We developed guides to answer each of these questions that either provided the most obviously "positive" answer or answered randomly from a bank of responses. Our applicants always answered "No" to any questions about possessing a prior criminal record.

Job sampling

We developed code that scraped all vacancies posted on each firm's proprietary hiring portal each day. We then manually identified the set of job titles that did not require a) a bachelor's or advanced degree, b) substantial prior experience, or c) a specialized license (at the time of application). When adding a new job for each firm, we selected randomly from among the most recently posted vacancies in counties from which we had not previously sampled a job for that firm. In rare cases no jobs were available in counties we had not previously sampled. In these cases we added new jobs in the same county but at different establishments to those sampled previously.

The *RandRes* platform automatically monitored scraped vacancies and added new jobs to the system. In each wave, we randomly sorted firms and worked through the sample by adding 5-10 jobs for each firm at a time to match maximum total application submission capacity.

Resume creation

RandRes features a PDF generator program that randomizes layout and design features to produce realistic resumes submitted as part of our application packages. The program parses an applicant's information generated by *RandRes*, include demographic details, employment history, and education history, and then randomly assigns a resume format including margins, font, text size, alignment, bullet shape, and other typical features. The process may redraw some features to ensure that resumes do not exceed one page in length or contain excessive white space.

The order and method in which information is presented is also random, meaning some applicants may list their education first while others list work experience first. Some resumes may include a separate section for references while others may include it as part of their employment history. Variations in language, such as whether or not to abbreviate U.S. state names, are also randomized.

The program tracks indicators of which special design attributes which have already been used in resumes for previous applicants at a particular job. This includes attributes such as off-white background coloring or a border around the contact information. Some resumes included monograms and watermarks as special attributes. A given resume may incorporate several of these design attributes together, but each special attribute is not used more than once at each job to ensure resumes are sufficiently differentiated. We find no evidence that special resume features increase contact rates.

We used the PDF resumes to signal characteristics not always collected in the online job application, such as year of high school graduation. When an applicant was assigned an LGBTQ or other student-club, the resume listed the club as part of educational experience. When an applicant was assigned preferred pronouns, they were listed in the resume below the applicant's name.

Application submission

The *RandRes* application platform automatically generated applications for all jobs active in the system. Applications were generated in pairs and new applications were generated whenever a job had fewer than two unsubmitted applications and no applications submitted within 24 hours. During Wave 1 of the experiment, applications were manually submitted by our team of undergraduate volunteers. *RandRes* instructed each volunteer which application to submit, provided the relevant details, and recorded submission status.

In subsequent waves, we developed software to automatically submit our applications to firms' job portals. By controlling a web browser, the software was able to visit the portal, fill out all application details, submit the application, and complete any assessments while operating at speeds designed to mimic human behavior. We used cloud computing providers to cycle through hundreds of IP addresses, user-agent strings, and other browser signatures to minimize our chances of detection.

We submitted up to 8 total applications to each job. Occasionally, vacancies would be closed or removed from hiring portals partway through our application process. Ninety-four percent of applicants were sent in complete groups of 8 and 88% of jobs received all 8 applications.

Appendix C Covariates

This Appendix provides details on sources and construction for the covariates used in Section 8.

Establishment-level covariates

- % county Black: Sourced from the U.S. Census's Longitudinal Employer-Household Dynamics Workplace Area Characteristics series. Measures the Black share of workers in 2015-2017 in the target job's county.

- % block Black / female: Same as above but defined at the census block level. Exact address data are not available for all jobs, making it impossible to match all jobs to census blocks. Only matched jobs are included.

- County IAT: Constructed using raw data from Harvard's Project Implicit. Defined as the average of all valid 2015 - 2020 IAT scores in each county, normalized to have a standard deviation of one within year. A higher value indicates more implicit bias against Black or female faces in the test. The female IAT used contrasts male vs. female faces with Science vs. Liberal Arts.

- DMA animus: Relative Google search rates for racially charged epithets as studied in Stephens-Davidowitz (2014). DMA refers to the target job's Designated Market Area. Higher values indicate more racially charged searchers. Normalized to have a standardized deviation of 1 within year and averaged over 2015-2019.

- State animus: Same as above but defined at state-level.

- White manager: Sourced from Reference USA establishment-level data. White manager indicates that Reference USA listed at least one "Manager", "Site Manager", or "Office Manager" as ethnically "Western European", "Eastern European", "Scandinavian", or "Mediterranean." Not all establishments were able to be linked to the Reference USA data, and not all establishments in Reference USA had manager ethnicity information. Only jobs with valid data are included. Constructed with the most recently available Reference USA data set.

- Male manager: Same as above but defined as at least one manager with gender listed as "Male."

- Log employment: Sourced from Reference USA establishment-level data. Normalized to have standard deviation of one in sample.

Firm-level covariates. All firm-level covariates are normalized to have a standard deviation of one in sample.

- Log employment: Total US employment scraped from most recent publicly available data online, including annual reports and firm websites.

- DOL cases/emp: Includes all wage and hour compliance actions since FY 2005 reported by the Department of Labor. Normalized by total employment.

- Empl-discr cases/emp: Data scraped from `https://www.goodjobsfirst.org/violation-tracker`. Defined as the total count of reported penalties since 2000 where the primary offense category is "Employment Discrimination" divided by employment. Firms with no penalties reported are coded as zeros.

- Sales / emp: Data from Dun and Bradstreet. Defined as total sales divided by DB-reported employment averaged over 2010-2018.

- Profit / emp: Data from Compustat. Defined as average gross profit divided by Compustat-reported employment averaged over 2010-2018. Three firms do not have Compustat data and are omitted.

- % board Black: Measures the average Black share of the corporate board over 2014-2019. Board member race sourced from blackenterprise.com and manual searches.

- Chief diversity officer: Binary indicator manually scraped from company websites. Includes C-Suite executives only.

- GD score: Overall company rating scraped from GlassDoor.com.

- GD diversity score: Diversity score ratings scraped from GlassDoor.com.

- Callback centralization: Defined as total number of unique phone numbers that contacted applicants the firm divided by the total number of jobs where applicants received at least one contact times minus 1. To avoid any mechanical correlation with outcomes, constructed as a leave-out mean omitting any contacts to own job.

- % managers white: Sourced from Reference USA. Measures share of managers at all establishments belonging to this firm with race reported as defined in establishment-level covariates. Two firms do not appear in the Reference USA data.

- % managers male: Same as above but defined as share of managers reported to be male.

Industry-level covariates.

- White - Black adj wage, male - female adj wage: Constructed using the CPS Monthly Outgoing Rotation Groups from 2009 to 2019, extracted from IPUMS at `https://cps.ipums.org/cps/`. Sample includes individuals aged 20-60 who work full-time (35+ hours a week) in the private sector that do not have imputed earnings or hours worked. To obtain 2-digit SIC industry codes, we link IPUMS variable IND1990 with 1987 SIC industry codes using a crosswalk from Autor, Dorn, and Hanson (2019). Wage gaps are obtained from a regression of log hourly wages (equal to weekly earnings divided by usual hours worked per week) on indicators for each industry, for being black (female), their interaction, and a set of year indicators. Adjusted wage gaps are the coefficients on the interaction between black (female) and industry. All calculations use CPS household or earnings weights.

- % ind Black, % ind female: Constructed using the Equal Employment Opportunity Commission's 2018 public use file of EEO-1 data. Defined as the Black (female) share of workers in the NAICS 3-digit industry.
- % mgmt - % ind Black, % mgmt - % ind female: Constructed using same data as above. Defined as the Black (female) share of mid-level officers and managers less the total Black (female) share of workers in the NAICS 3-digit industry.
- White - Black col share, male - female col share: Constructed using the same CPS sample and data as adjusted wage gaps. College share gaps are equal to the Black-white difference in the share of workers with a college degree in each industry.
- Top 4 sales share: Defined as the share of total sales accounted for by the four largest firms at the NAICS 3-digit level. Sourced from 2017 Economic Census data.

Occupation-level covariates.

- O*NET occupation task measures: We follow Deming (2017) and use the Occupational Information Network (O*NET), available at `https://www.onetcenter.org/db_releases.html`, to measure characteristics of occupations in the U.S.[28] The O*NET database provides information on various components of an occupation, including the *skills*, *knowledge*, and *abilities* required to perform the work, the *activities* typically performed on the job, and the *context*, or characteristics and conditions, of the job. We use this information to create the following five composite variables:

  - Analytical: Our analytic measure combines the following three components: 1) *mathematical reasoning ability* (defined as "the ability to understand and organize a problem and then to select a mathematical method or formula to solve the problem"), 2) *mathematics knowledge* ("knowledge of numbers, their operations, and interrelationships including arithmetic, algebra, geometry, calculus, statistics, and their applications"), and 3) *mathematics skill* ("using mathematics to solve problems").

[28] Unlike Deming (2017), we use production release 25.3 of O*NET.

- Social: Our social measure combines the following three skills: 1) *social perceptiveness* (defined as “being aware of others’ reactions and understanding why they react the way they do”), 2) *coordination* (“adjusting actions in relation to others’ actions”), 3) *persuasion* (“persuading others to approach things differently”), and 4) *negotiation* (“bringing others together and trying to reconcile differences”).
- Routine: Our routine measure combines two context variables, in particular 1) degree of automation (defined as “the level of automation of this job”) and 2) importance of repeating same tasks (“how important is repeating the same physical activities or mental activities over and over, without stopping, to performing this job?”).
- Service: Our service measure measure combines the activity variable *assisting and caring for others* (defined as “providing assistance or personal care to others”) and the skill variable *service orientation* (“actively looking for ways to help people”).
- Manual: Our manual measure combines two skill variables, specifically 1) *performing general physical activities* (defined as “performing physical activities that require considerable use of your arms and legs and moving your whole body, such as climbing, lifting, balancing, walking, stooping, and handling of materials”) and 2) *handling and moving objects* (“using hands and arms in handling, installing, positioning, and moving materials, and manipulating things”).
- Customer interaction: Our customer interaction measure averages two activities variables, one knowledge variable, and one context variable. The work activities variables include 1) *performing for or working directly with the public* (defined as “performing for people or dealing directly with the public”) and 2) *establishing and maintaining interpersonal relationships* (“developing constructive and cooperative working relationships with others, and maintaining them over time”). We use the work knowledge variable *customer and personal service* (“knowledge of principles and processes for providing customer and personal services) and the work context variable *contact with others*, which answers the question “how much does this job require the worker to be in contact with others (face-to-face, by telephone, or otherwise) in order to perform it?”

Each composite variable is calculated as the average of its component variables. Since some of these component variables are measured on different scales, we first rescale all the component variables to fall between 0 and 10.

Appendix D Technical Appendix

Denote the realized contact gap at job $j \in \{1, ..., J_f\}$ of firm $f \in \{1, ..., F\}$ by $\hat{\Delta}_{fj}$. For most of our analysis $\hat{\Delta}_{fj}$ is measured as the difference between white and Black contact rates at job j, but the same construction is used to study other binary protected characteristics such as gender. Let $D_{fj} \in \Omega$ give the *design* (i.e., assigned characteristics) of the portfolio of resumes sent to job j. This design includes, for example, the mix of employment histories on each resume, the time of day each resume was sent, each applicant's year of high school graduation, and the formatting of the resumes. Define $\hat{\Delta}_{fj}(d)$ as the contact gap that would arise at job j if it had been assigned application design d. Realized contact gaps can be written $\hat{\Delta}_{fj} = \hat{\Delta}_{fj}(D_{fj})$. Population contact gaps are defined as

$$\Delta_{fj} \equiv \mathbb{E}\left[\hat{\Delta}_{fj}(D_{fj}) \mid \left\{\hat{\Delta}_{fj}(d)\right\}_{d\in\Omega}\right] = \sum_{d\in\Omega} \omega_{fjd}\hat{\Delta}_{fj}(d),$$

where $\omega_{fjd} \in (0, 1)$ is the probability that design d is assigned to job j of firm f. Note that the expression after the equals sign presumes that the assignment probabilities $\{\omega_{fjd}\}$ are independent of the potential contact gaps $\{\hat{\Delta}_{fj}(d)\}$, a property ensured by random assignment. Assignment probabilities may differ by f as, for example, applicant job histories were tailored to the firms being studied. The $\{\omega_{fjd}\}$ may also differ across jobs, as local educational institutions and references were listed on applicant resumes.

We now make two key assumptions:

Assumption 1 (Design uncertainty) *The errors* $\left\{\hat{\Delta}_{fj} - \Delta_{fj}\right\}_{f=1,j=1}^{F,J_f}$ *are mutually independent and have mean zero.*

Assumption 2 (Sampling uncertainty) *Each firm's population gaps* $\{\Delta_{fj}\}_{j=1}^{J_f}$ *are iid draws from a firm specific distribution* G_f *with mean* Δ_f*.*

Assumption 1 follows from random assignment of application characteristics. This condition also implicitly requires the behavioral assumption of no interference between jobs, an assumption made more plausible by the requirement that sampled jobs be located in different U.S. counties. Assumption 2 follows from *i.i.d.* sampling of jobs from the set of available vacancies posted on company job boards. The mean Δ_f, which is our measure of discrimination at firm f, gives the expected contact gap at an average job posting by firm f over the course of our study.

Together, these assumptions yield a hierarchical model with two sources of uncertainty. The first source ("design uncertainty") arises from randomness in the application design assigned to each job. The second ("sampling uncertainty") arises from randomness in the set of jobs sampled. We use the operator $\mathbb{E}[\cdot]$ to denote expectations with respect to both sorts of uncertainty; that is, to denote integration against G_f and the design probabilities $\{\omega_{fjd}\}_{d\in\Omega}$. Our assumptions thus far imply that

$$\mathbb{E}\left[\hat{\Delta}_{fj}|\Delta_{fj}\right] = \Delta_{fj}, \quad \mathbb{E}\left[\hat{\Delta}_{fj}\right] = \Delta_f.$$

Target parameter

The variance of the firm component of discrimination can be defined as

$$\begin{aligned}\theta &= \frac{1}{F}\sum_{f=1}^{F}\Delta_f^2 - \left(\frac{1}{F}\sum_{f=1}^{F}\Delta_f\right)^2 \\ &= \left(\frac{F-1}{F}\right)\left\{\frac{1}{F}\sum_{f=1}^{F}\Delta_f^2 - \frac{2}{F\left(F-1\right)}\sum_{f=2}^{F}\sum_{k=1}^{f-1}\Delta_f\Delta_k\right\}.\end{aligned}$$

Bias corrected estimator

The fundamental difficulty in estimating θ involves the first term in the curly brackets. Let $\hat{\Delta}_f = \frac{1}{J_f}\sum_{j=1}^{J_f}\hat{\Delta}_{fj}$ denote the mean contact gap at firm f. Both design and sampling uncertainty generate an upward bias in the "plug-in" estimator $\left(\hat{\Delta}_f\right)^2$ of Δ_f^2 because

$$\begin{aligned}\mathbb{E}\left[\left(\hat{\Delta}_f\right)^2\right] &= \mathbb{E}\left[\left(\hat{\Delta}_f - \Delta_f\right)^2\right] + \Delta_f^2 \\ &= \mathbb{E}\left[\left(\underbrace{\hat{\Delta}_f - \frac{1}{J_f}\sum_{j=1}^{J_f}\Delta_{fj}}_{\text{design error}} + \underbrace{\frac{1}{J_f}\sum_{j=1}^{J_f}\Delta_{fj} - \Delta_f}_{\text{sampling error}}\right)^2\right] + \Delta_f^2 \\ &> \Delta_f^2.\end{aligned}$$

The bias corrected estimator of θ is motivated by the approximation $\mathbb{E}\left[\left(\hat{\Delta}_f - \Delta_f\right)^2\right] \approx s_f^2$, where s_f is an estimated standard error. When this approximation holds exactly, we have $\mathbb{E}\left[\hat{\Delta}_f^2\right] = \Delta_f^2 + s_f^2$. The bias corrected estimator can be written

$$\begin{aligned}\hat{\theta} &= \left(\frac{F-1}{F}\right)\left\{\underbrace{\frac{1}{F-1}\sum_{f=1}^{F}\left(\hat{\Delta}_f - \frac{1}{F}\sum_{k=1}^{F}\hat{\Delta}_k\right)^2}_{\text{plug-in}} - \underbrace{\frac{1}{F}\sum_{f=1}^{F}s_f^2}_{\text{correction}}\right\} \\ &= \left(\frac{F-1}{F}\right)\left\{\frac{1}{F}\sum_{f=1}^{F}\left(\hat{\Delta}_f^2 - s_f^2\right) - \frac{2}{F\left(F-1\right)}\sum_{f=2}^{F}\sum_{k=1}^{f-1}\hat{\Delta}_f\hat{\Delta}_k\right\}.\end{aligned}$$

Variants of this estimator have been applied in several literatures (e.g., Krueger and Summers, 1998; Aaronson et al., 2007), though typically without the adjustment factor of $\frac{F-1}{F}$.

In our analysis, we employ the following standard error estimator

$$s_f = \sqrt{\frac{1}{J_f\left(J_f-1\right)}\sum_{j=1}^{J_f}\left(\hat{\Delta}_{fj}-\hat{\Delta}_f\right)^2}.$$

With this choice of s_f, $\hat{\theta}$ becomes an unbiased leave out variance component estimator of the sort proposed by Kline, Saggio and Sølvsten (2020). In particular, it can be shown that

$$\hat{\Delta}_f^2 - s_f^2 = \frac{2}{J_f\left(J_f-1\right)}\sum_{j=2}^{J_f}\sum_{\ell=1}^{j-1}\hat{\Delta}_{fj}\hat{\Delta}_{f\ell} = \frac{1}{J_f}\sum_{j=1}^{J_f}\hat{\Delta}_{f(j)}\hat{\Delta}_{fj},$$

where $\hat{\Delta}_{f(j)} = \frac{1}{J_f-1}\sum_{\ell\neq j}\hat{\Delta}_{f\ell}$ is the leave-job out mean contact gap at firm f.

Independence of the errors across jobs guarantees that $\mathbb{E}[\hat{\Delta}_{fj}\hat{\Delta}_{f\ell}] = \mathbb{E}[\Delta_{fj}]\mathbb{E}[\Delta_{f\ell}] = \Delta_f^2$, with the second equality following from random sampling of jobs (Assumption 2). Likewise, independence of both design and sampling errors across firms ensures that $\mathbb{E}[\hat{\Delta}_f\hat{\Delta}_k] = \mathbb{E}[\hat{\Delta}_f]\mathbb{E}[\hat{\Delta}_k] = \Delta_f\Delta_k$. Consequently, $\mathbb{E}[\hat{\theta}] = \theta$. Lemma 3 of Kline, Saggio and Sølvsten (2020) establishes consistency of $\hat{\theta}$ for θ as the total number of jobs $\sum_{f=1}^F J_f$ grows large. Asymptotic normality of $\hat{\theta}$ follows from Theorem 2 of Kline, Saggio and Sølvsten (2020).

Cross-wave estimator

The cross wave estimator of θ is analogous to $\hat{\theta}$ but uses cross-products of wave level, as opposed to job-level, average gaps to estimate Δ_f^2. Suppose that for any two waves $(\tau_1,\tau_2)\in\{1,...,T_f\}^2$

$$\mathbb{E}\left[\hat{\hat{\Delta}}_{f\tau_1}\hat{\hat{\Delta}}_{f\tau_2}\right] = \Delta_f^2 \quad \text{if } \tau_1\neq\tau_2,$$

where $\hat{\hat{\Delta}}_{f\tau}$ is the mean gap in wave τ. This moment condition would follow from Assumptions # 1 and # 2 if each firm's distribution of population job gaps were restricted to be time invariant. An unbiased estimator of Δ_f^2 is the (job-weighted) cross-wave analogue of this moment condition:

$$\widehat{\Delta_f^2} \equiv \frac{\sum_{\tau_1=2}^{T_f}\sum_{\tau_2=1}^{\tau_1-1} n_{f\tau_1}n_{f\tau_2}\hat{\Delta}_{f\tau_1}\hat{\Delta}_{f\tau_2}}{\sum_{\tau_1=2}^{T_f}\sum_{\tau_2=1}^{\tau_1-1} n_{f\tau_1}n_{f\tau_2}},$$

where $n_{f\tau}$ gives the number of jobs sampled from firm f in wave τ. Our corresponding unbiased cross-wave estimator of θ is

$$\left(\frac{F-1}{F}\right)\left\{\frac{1}{F}\sum_{f}\widehat{\Delta_f^2}-\frac{2}{F\,(F-1)}\sum_{f=2}^{F}\sum_{k=1}^{f-1}\hat{\Delta}_f\hat{\Delta}_k\right\}.$$

Cross-state estimator

The cross state estimator is identical to the cross-wave estimator except that cross-products between state averages of job contact gaps at each firm replace wave averages of job contact gaps at each firm. As with the cross-wave estimator, the cross-products of averages are job weighted.

Industry and portal intermediary variance components

Firm identifiers are "nested" within industry and job portal intermediary categories. Variance components for these alternate groupings of jobs can be defined as weighted analogues of the firm level component θ.

Working with industry as our focal example, let $\ddot{\Delta}_i$ denote the population contact gap in industry $i \in \{1,...,I\}$, which we define as the equally weighted average of the population contact gaps among firms in that industry. Letting F_i be the number of firms in industry i and $F=\sum_{i=1}^{I}F_i$ the total number of firms in the experiment, the industry component can be written:

$$\begin{aligned}\theta_I &= \frac{1}{F}\sum_{i=1}^{I}F_i\ddot{\Delta}_i^2-\left(\frac{1}{F}\sum_{i=1}^{I}F_i\ddot{\Delta}_i\right)^2\\ &= \left(\frac{F-1}{F}\right)\left\{\frac{1}{F(F-1)}\sum_{i=1}^{I}F_i(F-F_i)\ddot{\Delta}_i^2-\frac{2}{F\,(F-1)}\sum_{i=2}^{I}\sum_{k=1}^{i-1}F_iF_k\ddot{\Delta}_i\ddot{\Delta}_k\right\}\end{aligned}$$

The firm weighting used in this definition ensures that the ratio $\theta_I/\theta \in [0,1]$ possesses an R^2 interpretation. When $\theta_I=\theta$ industry explains all of the variation across firms.

Mirroring the firm-level analysis, an unbiased estimate of the squared mean $\ddot{\Delta}_i^2$ can be constructed as a weighted average of cross-products of job-level gaps in industry i. To preserve the interpretation of $\ddot{\Delta}_i$ as an equally weighted average of contact gaps across firms in an industry, we weight jobs inversely by "firm size" when computing these cross-products. Indexing jobs in industry i by $n \in \{1,...,N_i\}$, let $\hat{\Delta}_{in}$ give the estimated contact gap at that job. Using $f(i,n)$ to denote the parent company of job n our job weights can be written $w_{in}=1/J_{f(i,n)}$. Note that w_{in} gives the inverse of the total number of jobs at the parent firm containing job n. Hence, an unbiased estimator of $\ddot{\Delta}_i$

is $\left(\sum_{n=1}^{N_i} w_{in}\right)^{-1}\left(\sum_{n=1}^{N_i} w_{in}\hat{\Delta}_{in}\right)$. Our corresponding estimator for $\ddot{\Delta}_i^2$ can be written:

$$\widehat{\ddot{\Delta}_i^2} \equiv \frac{\sum_{n=2}^{N_i}\sum_{k=1}^{n-1} w_{in}w_{ik}\widehat{\Delta}_{in}\widehat{\Delta}_{ik}}{\sum_{n=2}^{N_i}\sum_{k=1}^{n-1} w_{in}w_{ik}}.$$

Plugging these unbiased estimators of $\ddot{\Delta}_i$ and $\ddot{\Delta}_i^2$ into the expression for θ_I yields the unbiased industry variance component estimator $\hat{\theta}_I$.

State and job title variance components

Defining state and job title variance components requires some additional notation, as these groupings of jobs do not nest firms. Working with state as our focal example, we index states by $s \in \{1, ..., S\}$ and jobs in states by $b \in \{1, ..., B_s\}$. Accordingly, we denote the population gap at job b of state s by Δ_{sb}. Letting $w_{f(s,b)} = 1/J_f$ denote the inverse size of the firm containing job b, and $W_s = \sum_{b=1}^{B_s} w_{f(s,b)}$, the sum of these weights, the overall population gap in state s is defined as

$$\dddot{\Delta}_s = \frac{1}{W_s}\sum_{b=1}^{B_s} w_{f(s,b)}\Delta_{sb}.$$

Letting $W = \sum_{s=1}^{S} W_s$ be the total number of firms in the experiment, our variance component of interest is:

$$\begin{aligned}
\theta_S &= \frac{1}{W}\sum_{s=1}^{S} W_s\dddot{\Delta}_s^2 - \left(\frac{1}{W}\sum_{s=1}^{S} W_s\dddot{\Delta}_s\right)^2 \\
&= \left(\frac{W-1}{W}\right)\left\{\frac{1}{W(W-1)}\sum_{s=1}^{S} W_s(W-W_s)\dddot{\Delta}_s^2 - \frac{2}{W\,(W-1)}\sum_{s=2}^{S}\sum_{k=1}^{s-1} W_sW_k\dddot{\Delta}_s\dddot{\Delta}_k\right\}.
\end{aligned}$$

To estimate θ_S we substitute $\widehat{\dddot{\Delta}}_s = \frac{1}{W_s}\sum_{b=1}^{B_s} w_{f(s,b)}\hat{\Delta}_{sb}$ for $\dddot{\Delta}_s$ in the second term in braces. The quantity $\dddot{\Delta}_s^2$ entering the first term in braces is replaced with the weighted average cross-product:

$$\frac{\sum_{b=2}^{B_s}\sum_{k=1}^{b-1} w_{f(s,b)}w_{f(s,k)}\widehat{\Delta}_{sb}\widehat{\Delta}_{sk}}{\sum_{b=2}^{B_s}\sum_{k=1}^{b-1} w_{f(s,b)}w_{f(s,k)}}.$$