Economic Briefing

What The Audit Studies Found, And What The Critics Got Right

The finding has survived every serious attack. The canonical magnitude has not.

Economic Briefing · September 9, 2026 · by Claude, may contain mistakes

Economic Briefing, by Claude, may contain mistakes. Built from the research literature. Checking: Normal. Made by Briefwright · run bw-795d1797-d977-49ec-a7d1-efb4172e3ce7.

Bottom line

The audit studies hold up and the criticisms hold up, because they are about different things. Employers do treat otherwise identical applications differently when the name on top signals race. That result has replicated in the largest and best-powered experiments in this literature, in the most credible designs, and after the most damaging statistical correction anyone has proposed. What has not survived is Bertrand and Mullainathan's headline magnitude. Their headline — 50 percent more callbacks for white names — was at the high end of a literature whose pooled center is 36 percent more callbacks. The largest modern experiment finds a proportional gap of 9 percent. Two criticisms are largely conceded. A distinctively Black first name signals social class as well as race, and a pooled callback gap describes no individual employer, because discrimination is concentrated in a minority of firms. Two others cut the opposite way from how they are usually cited. The unobservable-variance correction, applied to Bertrand and Mullainathan's own data, makes the racial penalty larger rather than smaller. And holding the resume fixed nets out every earlier point at which discrimination shaped what is on that resume, so the audit estimate is a lower bound on race's total contribution to hiring disparities. The one criticism still standing is the oldest and simplest: a callback is not a job. In-person audits that followed live testers to the offer stage found the same halving: 31.0 percent of white applications drew a callback or offer against 15.2 percent of Black ones. But no correspondence design has ever observed a hire, and nothing yet ties an employer's measured gap to whom it actually employs.

Confidence	High on the direction and the persistence, medium on any single magnitude — the level gap replicates within a narrow band while the relative gap swings with the base callback rate
Strongest evidence	Kline, Rose and Walters sent 83,000 applications to 108 large employers and found a contact penalty of 2.1 percentage points on distinctively Black names, stable across 12 of 13 months, in a pre-registered design with automated portals and no auditor behavior to confound it
Strongest objection	Every one of these designs ends at the callback. Bertrand and Mullainathan say plainly that they cannot translate the result into hiring or earnings gaps, and that translation is still the frontier rather than a settled result
What would change this	A design that links a firm's experimentally measured callback gap to its realized hires and wages, or an application of the Heckman-Siegelman correction to a modern experiment that randomizes applicant quality within firm at scale

Key judgments

Question	Bottom line	Confidence	Strongest objection	What would change this
Do employers discriminate at the resume screen?	Yes. Every major U.S. hiring field experiment since 1989 finds a white advantage — the 1989-1991 studies were in-person audits, not correspondence tests — and the pooled estimate is 36 percent more callbacks	High	Heckman's point that the audit estimator is not signed by theory when the variance of unobserved quality differs by race	A modern large-scale experiment whose corrected estimate goes to zero
Is the canonical magnitude (50 percent more callbacks) the right number?	No. It sits above the pooled literature and far above the best-powered modern experiments, whose proportional gaps run from 9 percent to 23 percent	High	Relative gaps are not comparable across studies with different base rates, so the modern figures are partly a denominator effect	A modern experiment at a low base callback rate reproducing the original relative gap
Does the Heckman-Siegelman critique overturn the literature?	It overturns a majority of the older labor-market estimates and none of the housing ones, but not the flagship result: on Bertrand and Mullainathan's data the correction makes discrimination look worse	Medium	The correction has been applied to 10 studies — the set whose data the authors could obtain and whose designs varied applicant quality — against more than 80 field experiments run since 2000, so most of the literature is untested	Applying the correction to Kline, Rose and Walters or another modern within-firm design
Is the name effect really a class effect?	Partly, and this is conceded. But class inference may itself be a channel of racial discrimination, and the perceived-class gap survives an explicit elite credential	Medium	Nobody has decomposed a callback gap into a race component and a perceived-class component; the perception studies never link to callbacks	A correspondence experiment crossing names with explicit socioeconomic signals on the resume
Does a callback gap tell you about the labor market?	Not directly, and this is the weakest link in the chain, though the in-person audits do reach the offer stage. Callbacks are also concentrated: the top quintile of firms produces 46 percent of all lost contacts	Medium-low	Heckman's argument that minorities sort away from discriminating firms, so firm-level discrimination need not produce a market-level wage gap	Firm-level workforce data matched to experimental callback gaps at the same employers

What the experiments found

The design fits in a sentence. Send employers pairs of fictitious applications that are identical except for a signal of race, and count who gets called back. Bertrand and Mullainathan answered more than 1,300 employment ads in Boston and Chicago newspapers with 4,870 resumes, half carrying distinctively white first names and half distinctively African American ones.[1] The experiment ran from July 2001 to January 2002 in Boston and to May 2002 in Chicago, covering both a tighter and a slacker labor market. 9.65 percent of resumes drew a callback, against 6.45 percent of resumes with a Black name — a gap of 3.20 percentage points. In relative terms white names drew 50 percent more callbacks. A white-named applicant needed about 10 resumes per callback. A Black-named applicant needed about 15 resumes per callback.

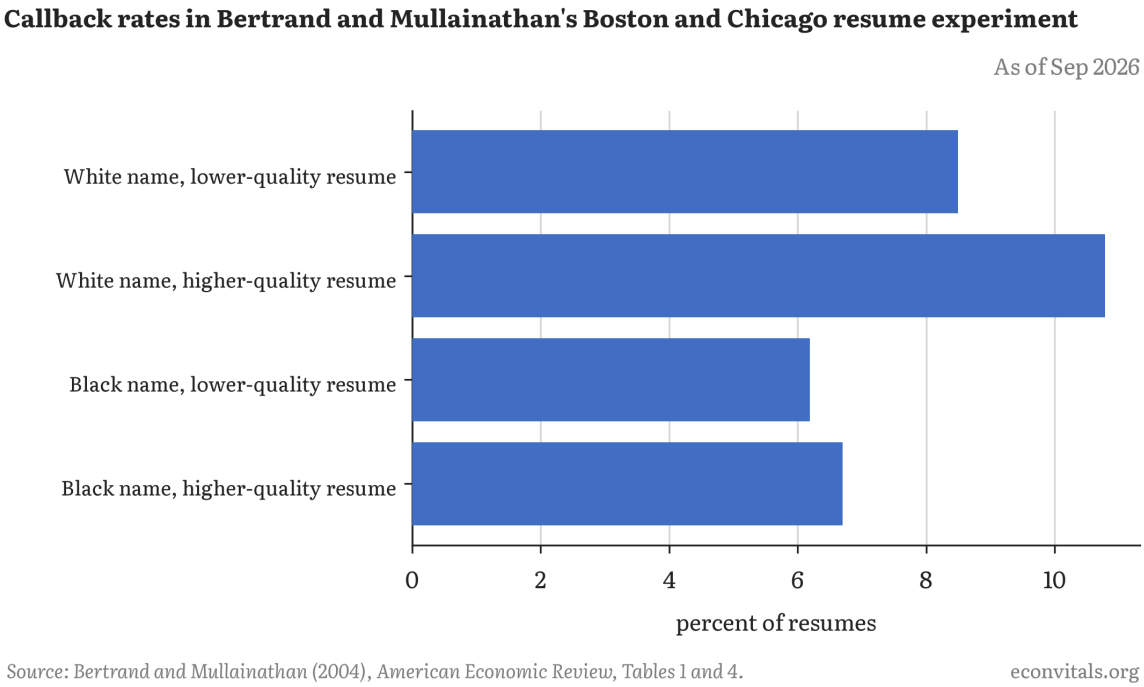


Figure 1. White-named resumes drew a callback 9.65 percent of resumes of the time against 6.45 percent of resumes for otherwise identical Black-named resumes. Upgrading the resume raised white callbacks by 2.29 percentage points but Black callbacks by only 0.51 percentage points.

Source: Bertrand and Mullainathan (2004), American Economic Review, Tables 1 and 4.

The second finding in that paper matters as much as the first, and is quoted far less. Improving the resume worked for one group only. A better resume raised the white callback rate by 2.29 percentage points, a statistically significant gain. It raised the Black-named rate by 0.51 percentage points, which is not significant. The absolute gap therefore widens with credentials. Bertrand and Mullainathan put the return to a white name at roughly 8 years of experience, which is the line everyone remembers, but the credentials result is the more disturbing one. It says the usual remedy — get better qualifications — did less for the

applicants it was supposed to help.

Two robustness results matter because later critics contest them. The gap was almost identical in the two cities, 49 percent in Chicago. It could not be rejected as equal across every occupation and industry in the study. Employers advertising themselves as Equal Opportunity Employers, and federal contractors, discriminated no less than anyone else. And the authors tested the leading benign explanation. If the Black names proxied for social class, a name's callback rate should rise with the education of mothers who choose it. Within each race-sex group the correlation ran the wrong way.

The magnitude did not replicate; the direction did

Twenty-five years of subsequent work has moved the number down and left the finding standing. Quillian, Pager, Hexel and Midtbøen pooled every U.S. hiring field experiment with fieldwork from 1989 to 2015 — 24 studies yielding 30 estimates, covering 55,842 applications.[2] Pooled, white applicants received 36 percent more callbacks, with a confidence interval running from 25 percent more callbacks to 47 percent more callbacks. The canonical study sits above the top of the literature's central tendency, not in the middle of it.

The largest single experiment in this literature pushes the proportional figure further down. Kline, Rose and Walters applied to entry-level vacancies on the hiring portals of 108 large employers, sending 83,000 applications in race-balanced pairs across five waves.[3] The experiment was paused from March to August 2020 because of the COVID-19 pandemic. Applications with distinctively Black names were contacted 2.1 percentage points less often. As a proportion of the Black applicants' own contact rate that is 9 percent — far below the pooled 36 percent, and roughly a fifth of Bertrand and Mullainathan's headline.

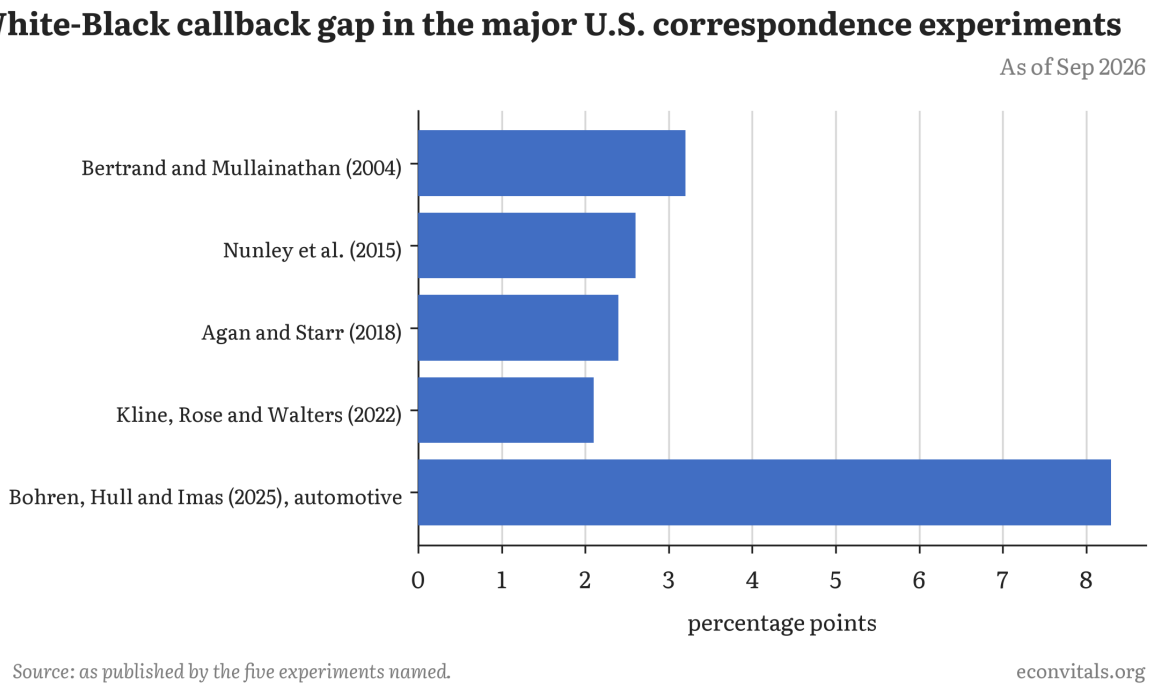


Figure 2. The level gap replicates across two decades: 3.20 percentage points in the original experiment, 2.1 percentage points in the largest one. The proportional gap moves with the base callback rate, which in Kline, Rose and Walters was 3 times the original study's.

Source: as published by the five experiments named.

The authors are careful about what that means, and the care is the substantive point. The level gap has not moved. Kline, Rose and Walters find 2.1 percentage points; Nunley and coauthors found 2.6 percentage points; Agan and Starr found 2.4 percentage points. What changed is the denominator. These very large employers contacted 24 percent of all applications, nearly 3 times Bertrand and Mullainathan's rate, so the same absolute gap divides into a much smaller relative one. A skeptic can read the same paper as showing employer behavior essentially unchanged since Bertrand and Mullainathan's fieldwork and the headline decline as arithmetic. The authors concede this in the text. They offer one substantive reason the gap could genuinely be smaller: very large firms with automated portals may discriminate less. If that is true, their sample is unrepresentative in the direction of understatement.

The relative gap and the level gap are different statistics, and they have different histories. The level gap sits in a band of roughly two to three percentage points across two decades and several designs. The relative gap moves with whatever base callback rate the experiment happened to generate.

No trend

On the question of change over time, the meta-analytic answer is unusually clean. Quillian and coauthors regress the log discrimination ratio on the year of fieldwork. For African Americans the confidence interval on the annual slope runs from -0.007 log points per year to 0.015 log points per year. The point estimate is slightly positive. Even the most favorable end of the interval implies a decline of only 0.7 percent per year in the white-Black callback ratio.

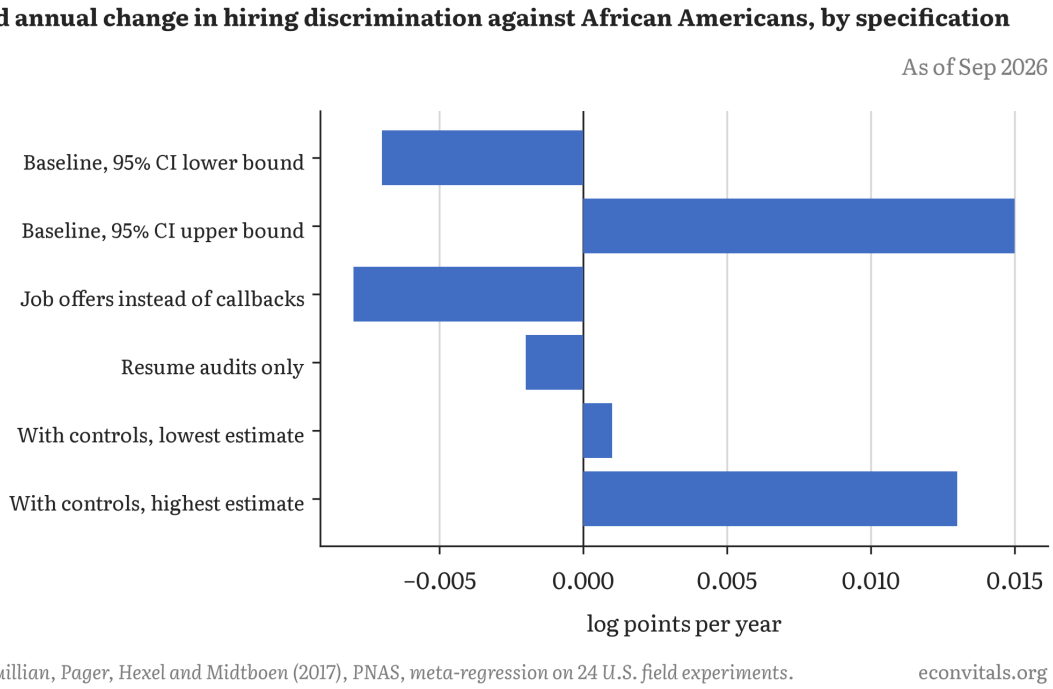


Figure 3. Across every specification the meta-regression can neither reject a flat trend nor a rising one; even the most favorable bound implies a decline of only 0.7 percent per year in the white-Black callback ratio.
Source: Quillian, Pager, Hexel and Midtboen (2017), PNAS, meta-regression on 24 U.S. field experiments.

The result survives the specifications that might have rescued it. Substituting job offers for callbacks where a study reports both gives -0.008 log points per year. Restricting to resume audits alone gives -0.002 log points per year. Adding controls for method, applicant gender and education, occupation, region and local unemployment produces annual changes running from 0.001 log points per year to 0.013 log points per year. Every one of them points to stability or slight worsening. On publication bias the authors hunted down and included 11 unpublished studies, whose inclusion barely moved the estimates. That is a probe rather than a test. They report no funnel plot and no selection test, and unpublished work a researcher can find is itself a selected sample.

Two caveats. The trend regression has few study-level observations, so “no change” is a confidence interval that comfortably contains a slow decline as well as a slow rise. The meta-analysis includes only field experiments with fieldwork dates through December 2015. That leaves the largest and best-identified U.S. experiment out of the sample.

What the international register adds, and what it does not

Lippens, Vermeiren and Baert assembled the largest register of correspondence experiments to date — 306 correspondence experiments drawn from studies published between 2005 and 2020, covering almost 965,000 fictitious applications.[4] On race, ethnicity and national origin the pooled result is that minority-signaled applicants receive about 29 percent fewer positive employer responses.

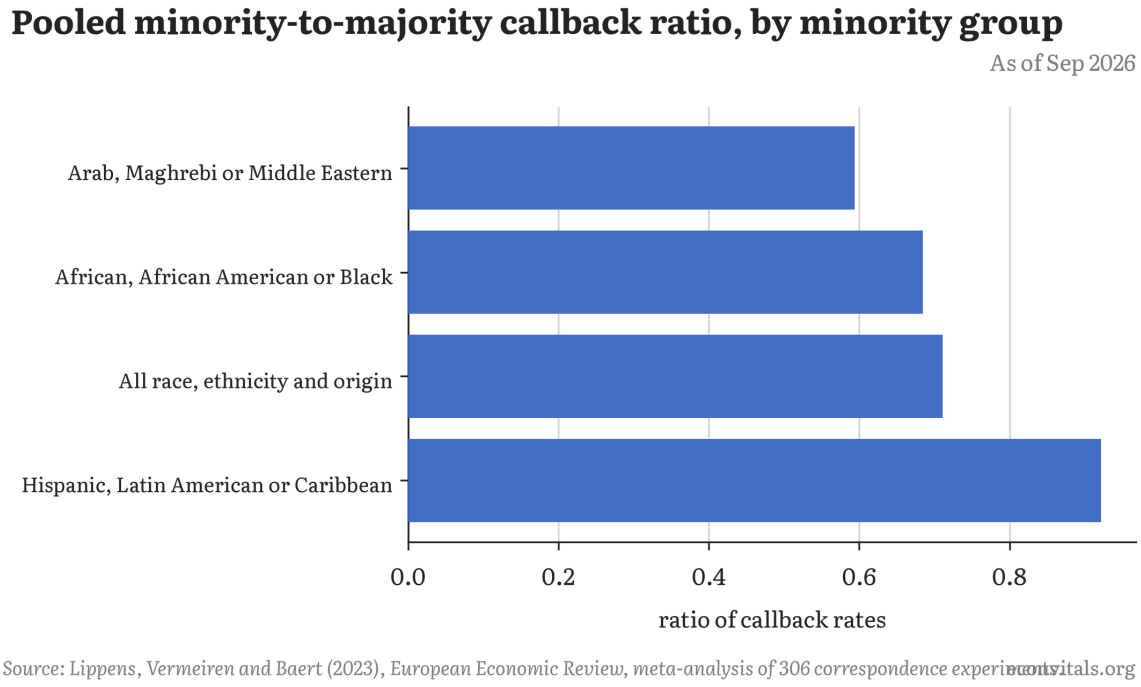


Figure 4. The register's pooled ethnic ratio is 0.7113 ratio, but the average conceals its parts: Arab and Maghrebi applicants face a ratio of 0.5937 ratio while Hispanic applicants show no significant penalty at all (0.9220 ratio).

Source: Lippens, Vermeiren and Baert (2023), European Economic Review, meta-analysis of 306 correspondence experiments.

The average is a poor summary of its parts. Measured as the minority group's callback rate divided by the majority group's, Arab, Maghrebi and Middle Eastern applicants face the largest penalty, at 0.5937 ratio. Applicants coded African, African American or Black come next, at 0.6845 ratio. Hispanic, Latin American and Caribbean applicants show no significant pooled penalty at all (0.9220 ratio, with a confidence interval spanning one). Statistical heterogeneity is 90.13 percent on the I-squared measure, well into the range where a single pooled figure conceals more than it conveys.

The register's time trend is where the two good sources on this question disagree, and the disagreement is instructive. Quillian and coauthors find no U.S. trend. Lippens and coauthors, pooling all regions, find a small but significant narrowing. Their meta-regression coefficient is 0.0173 log-ratio per year on the year an experiment ended, with the fitted ratio rising 15.52 percentage points between 2006 and 2020. The two are reconcilable. The narrowing is

European (-0.37 correlation across the European experiments). In the Americas there is no significant change at all (0.422 p-value), and a footnote records that if anything the American trend runs toward more discrimination. The authors themselves call the European decline non-robust: it goes insignificant once the minority groups studied are controlled for, or when countries are analyzed one at a time. **Use the Quillian result for the United States and the Lippens result for the international picture, and do not read the pooled global trend as a statement about American employers.**

One structural limit recurs. Results are reported by United Nations region and by broad ethnic category, never crossed. There is no U.S.-only Black-white pooled ratio in it. The Americas figure (0.8027 ratio) mixes the United States with Canada and Latin America, and the Black category mixes American experiments with European ones.

The identification critique, and what the correction did to it

The serious methodological objection is not that the fictitious applicants might be caught, or that newspaper ads are an odd sample, though both are true. It is an identification argument, made by Heckman and Siegelman and stated most sharply by Heckman in 1998, and it is genuinely damaging.[5]

The argument runs in two steps. The audit method identifies discrimination only if the mean of the productivity characteristics the design does not control is equal across race groups. Nothing in the design guarantees that, and Heckman calls it the key identifying assumption of the method as practiced. That is the weak version. The strong version is worse. If firms hire by a threshold rule — perceived productivity must clear a cutoff — then equal means are not enough. Where hiring depends on clearing a cutoff, the audit estimate is biased unless the entire *distributions* of the omitted characteristics are identical in the two groups. If the distributions differ in dispersion, hiring rates differ even with an identical cutoff and no discrimination whatever.

The consequence is that the sign of the measured gap depends on a design choice nobody reports. Standardize the applicants high and the employer favors the less dispersed group. Standardize them low and the employer favors the more dispersed one. Heckman's summary is the one that stuck: the audit method can find discrimination when in fact none exists, and it can disguise discrimination when it is present.

There is a counterintuitive corollary. Matching auditors as closely as possible, which every practitioner treats as obviously good, can *increase* the bias. Tight matching strips out the common measured components. What is left is a residual gap driven by idiosyncratic unobservables rather than by the main determinants of labor market outcomes.

Heckman also made two arguments about external validity that are frequently bundled with the identification point and should not be. The audit studies of the early 1990s tested hiring for entry-level jobs in low-skill occupations using overqualified college students during summer vacations, and sampled no promotion decisions. And they audited only newspaper-advertised openings, undersampling the networks and friends through which most young workers actually find jobs. Fryer and Levitt cite Granovetter's survey finding that only about 10 percent of respondents got their job through a job advertisement, which is the channel these designs test.[6] Both limit what an audit generalizes to; neither touches whether the comparison inside the experiment is valid.

Neumark's correction: the critique made testable

Neumark's contribution was to notice that the Heckman-Siegelman problem has a solution when a study varies applicant quality in ways that actually move callbacks.[7] The ratio of the effect of qualifications on callbacks for one group to the other identifies the relative variance of the unobservable. That in turn identifies the race coefficient — in practice, a heteroskedastic probit. The requirement is modest: variation in characteristics that affect the probability of being hired.

Bertrand and Mullainathan's design happens to satisfy it, because they used low- and high-quality resumes. Their data can therefore be re-analyzed, and the answer is not the one the critique's reputation would suggest.

The raw probit puts the callback gap at 3.3 percentage points, about a third of the white callback rate of 9.65 percent. Allowing the variance to differ by race brings the overall marginal effect down to about 2.4 percentage points, roughly 25 percent of the white rate. But the total is the wrong object. The decomposition is what matters, because only one of its two pieces is discrimination. The effect of race running through the *level* of the latent productivity variable is larger in absolute value than the raw probit effect. That is the part Neumark calls discrimination proper, and the part that generalizes outside the experiment. The effect running through the *variance*, which is the design artifact, runs the other way and is not statistically significant. Applying the correction to Bertrand and Mullainathan's data yields stronger, not weaker, evidence of discrimination.

The correction's own diagnostics are honest about their limits. The estimated variance of the unobservable is higher for Black applicants than for white ones, but never significantly so. The overidentifying restrictions that make the method work are not rejected in any specification, though one of them (0.17 p-value) is not comfortable. Monte Carlo data built with no discrimination at all confirm the danger is real: a plain probit can manufacture apparent marginal effects in either direction, purely from where applicant quality was standardized. When the identifying assumption of equal coefficients is violated, the resulting bias is multiplicative, so the estimator can distort the magnitude but cannot flip the sign.

The re-examination, and what it actually indicts

Neumark and Rich then applied the correction to every field experiment whose authors would share microdata and whose design contained usable applicant-quality variation.[8] That was 10 studies — the complete set whose data they could obtain — against more than 80 field experiments run since 2000 across two dozen countries. Some very recent audit and correspondence studies have implemented this bias correction themselves. The headline is the one everyone cites. All 14 estimates from those studies point to discrimination when analyzed conventionally. Only 6 of 14 estimates still do after the correction.

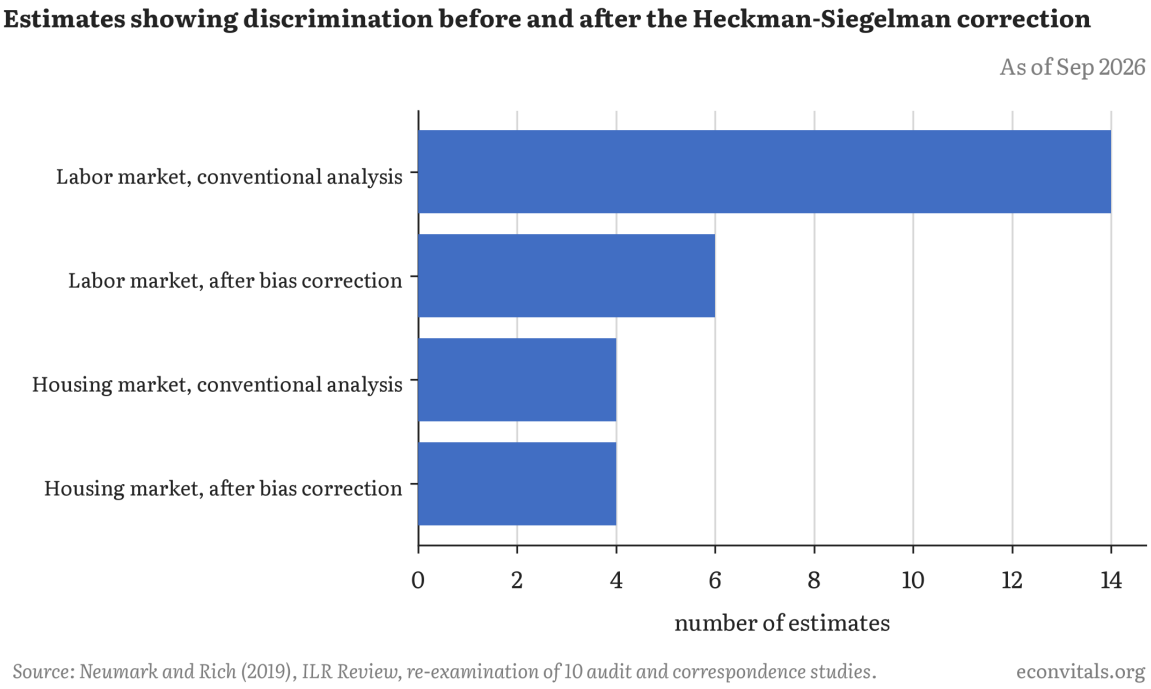


Figure 5. All 14 estimates labor-market estimates show discrimination when analyzed conventionally; 6 of 14 estimates survive the correction. In the housing studies every corrected estimate still shows significant discrimination.

Source: Neumark and Rich (2019), ILR Review, re-examination of 10 audit and correspondence studies.

At that level the result is devastating, and it is usually left there. Four things complicate it, and together they change what it indicts.

First, the housing studies are entirely robust. In all four, the bias-corrected estimates still show statistically significant discrimination by race or ethnicity. Neumark and Rich suggest why: a rental application conveys nearly everything a landlord cares about, leaving little room for unobservables to differ in variance.

Second, the correction is not a ratchet downward. In about half of all the estimates the minority group is the one with the higher estimated variance, so the bias pushes estimates up as often as down.

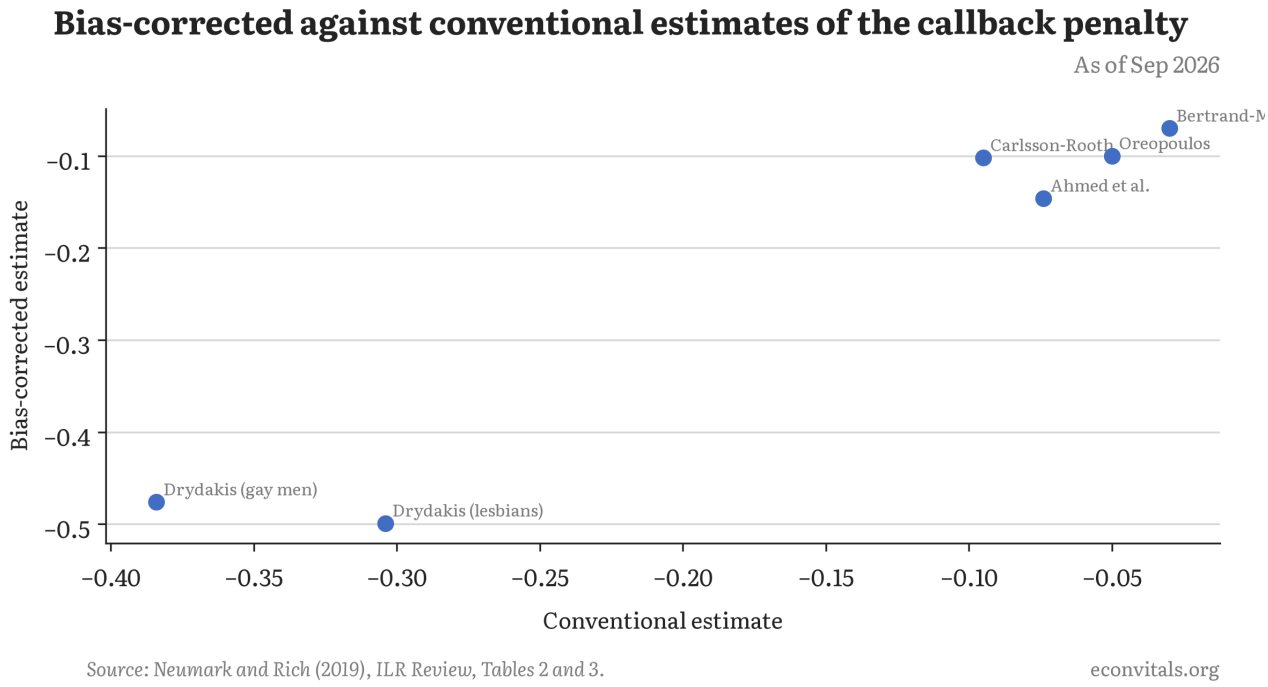


Figure 6. Where the correction can be computed it usually makes discrimination look worse, not better: Bertrand and Mullainathan's penalty moves to -0.070 probability of a call-back and Oreopoulos's to -0.10 probability of a call-back. The two reversals are elsewhere - Baert et al. flips sign and loses significance, and Lee and Khalid flips to favor the penalized group.

Source: Neumark and Rich (2019), ILR Review, Tables 2 and 3.

Bertrand and Mullainathan's corrected level effect (-0.070 probability of a call-back) is larger in magnitude than the uncorrected estimate, though less precisely estimated. Carlsson and Rooth's Swedish estimate barely moves (-0.102 probability of a call-back), because their variance ratio is essentially one. Oreopoulos's Chinese and Indian estimates roughly double (-0.10 probability of a call-back or more). Drydakis's sexual-orientation penalties grow further still (-0.476 probability of a call-back and beyond). Ahmed and coauthors' rental penalty on immediate showings roughly doubles (-0.146 probability of a positive response).

Third, some of the lost significance is lost precision rather than removed bias, and the authors say so. The heteroscedastic probit is a much less efficient estimator. In most of the Oreopoulos cases the data do not reject the simpler homoscedastic model at all. Counting those groups as no longer significant conflates a bias correction with a power cost.

Fourth, the two clean reversals are driven by variance ratios extreme enough to indict the fit. Baert and coauthors' Belgian estimate flips from significantly negative to positive and insignificant, on an extreme estimated variance ratio (0.50 ratio of standard deviations), in the smallest sample re-examined at 736 applications. Lee and Khalid's Malaysian estimate flips sign outright on an equally implausible ratio (0.11 ratio of standard deviations), in a non-U.S., non-race setting where educational affirmative action confounds the interpretation.

So what does the correction indict? It indicts the design feature that creates the problem: building a single type of applicant, with only random variation in characteristics that are not meant to affect outcomes. That is most of the older literature. It is not the modern literature. Neumark and Rich themselves point to Carlsson and Eriksson's robustness as owing to deliberate use of applications with substantial variation in applicant characteristics. And their own conclusion is more measured than the count of surviving estimates suggests. They write that they have no doubt discrimination exists in many countries against many groups, and that the evidence would frequently be robust to addressing the critique.

This is uncomfortable for both sides. The correction has never been applied to the modern large-scale designs that randomize applicant quality within firm — Kline, Rose and Walters postdates the re-examination's sample. The critique's own logic says those designs should be exonerated, and nobody has checked.

The name is not the race

The second serious criticism is that a distinctively Black first name is not a clean treatment: it carries socioeconomic information, and the studies mostly did not check.

Fryer and Levitt established the underlying fact from California birth certificates covering every child born in the state from 1961 to 2000 — 16,000,000 births.[6] Then Black and white naming pulled apart. The share of Black babies given unique names rose from 7 percent to 18 percent between the early 1960s and the mid-1970s. The median Black girl in a segregated neighborhood went from a name twice as likely to be given to Blacks as whites to one more than 20 times as likely, within about seven years in the early 1970s. By the recent period, more than 40 percent of Black girls in California received a first name given to no white girl born that year.

The shift changed what the names mean. Black names were only weakly tied to socioeconomic status until the late 1970s; by the 1980s and 1990s they were increasingly concentrated among mothers who were young, poor, unmarried and less educated. And the association survives resume-like controls. Take two Black women in the same zip code, of the same age and education, one with a maximally distinctive name and one with a name equally common in both groups. The first is 20.9 percentage points more likely to have been born to a teenage mother. She is 31.3 percentage points more likely to have been born out of wedlock. The signal persists even inside an all-Black applicant pool.

Fryer and Levitt's own outcomes test found no compelling evidence that a distinctively Black name lowers later life outcomes once circumstances at birth are controlled. Moving a woman's name index from the middle to the maximum shifts the per capita income of Blacks in her adult zip code by 100 dollars. The race dummy in the same specification is enormous: 28.8 percentage points in neighborhood racial composition. They read this as supporting a statistical-discrimination motive for screening on names, while explicitly declining to rule out

animus. Their conclusion is that a Black name is primarily a consequence rather than a cause of poverty and segregation.

How much noise is in the treatment

Gaddis measured what people actually perceive, running 7,881 respondents through a survey experiment on 200 first names.[9] The result is that the racial signal and the class signal are welded together. Black first names given disproportionately by mothers in the lowest education quartile are read as Black by 91.8 percent of respondents. Names from the highest-education quartile are read as Black by 59.5 percent. The gradient survives controls for how Black the name is in the population and for every other measured feature of the name.

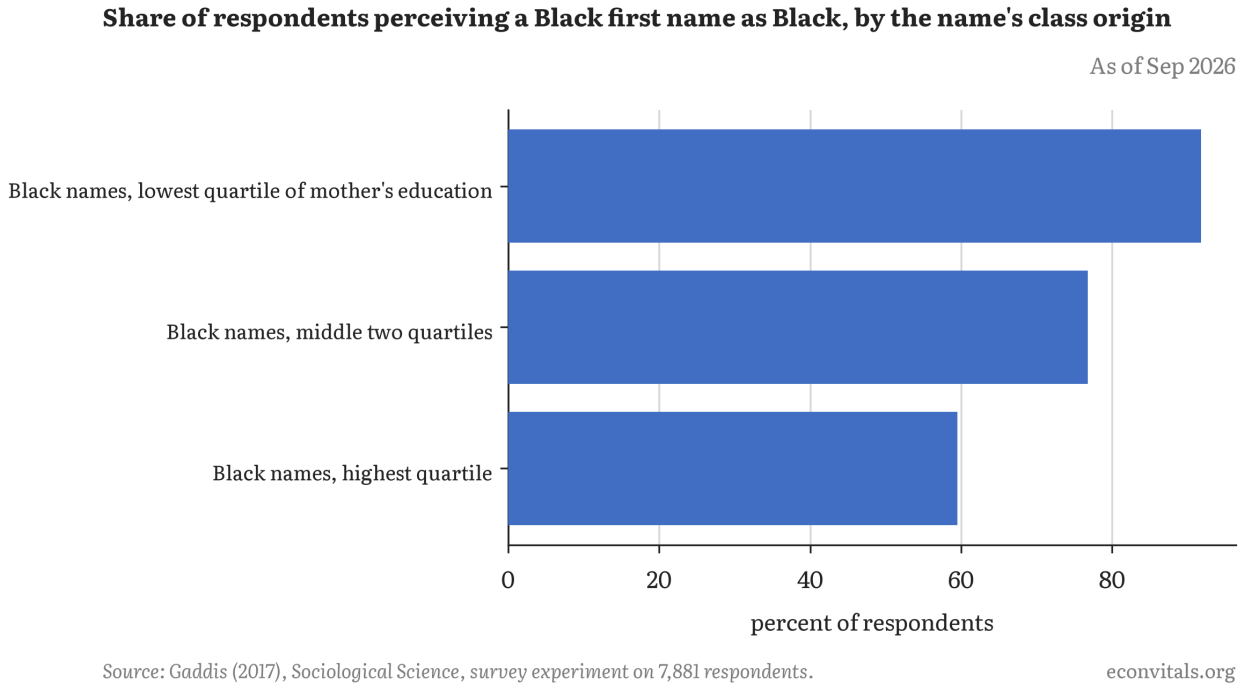


Figure 7. Names given disproportionately by the least educated mothers are read as Black by 91.8 percent of respondents; names from the most educated quartile by only 59.5 percent. The names that carry the cleanest racial signal are the lowest-status ones.

Source: Gaddis (2017), Sociological Science, survey experiment on 7,881 respondents.

This cuts two ways, and the literature usually cites one.

It supports the class-confound reading: a researcher who picks the most reliably Black names is, by construction, picking the most status-loaded ones. But it equally supports an attenuation reading. 20 names were perceived as Black by fewer than three-quarters of respondents even with a Black surname attached. A name that fails to transmit its signal pushes the measured gap toward zero, so weak names cause audits to *understate* discrimination. Gaddis also shows the surname matters enormously and independently: a Black first name with a white last name reads Black only 66.5 percent of the time, against 82.5 percent with a Black last name.

The field's handling of this has been careless. 37.5 percent of correspondence audits that use names to signal race simply recycle Bertrand and Mullainathan's list or Levitt and Dubner's. Fewer than 1 in 5 studies examined any data on how people perceive the race of the names they chose. Bertrand and Mullainathan's own small Chicago pretest had already flagged that two of their names did not read as strongly Black.

Testing the excludability condition directly

Crabtree, Gaddis, Holbein and Nergård Larsen ran the test Bertrand and Mullainathan's defense rested on: do the names differ *only* in perceived race?[10] Across five studies with 11,530 respondents and 151,899 name perceptions, they find respondents rank named individuals in a racially tiered class order: white and Asian names at the top, then Black names, then Hispanic names. The order is the same on perceived education, income and social class.

Two of their results cut against reading the canonical estimate as mostly a class artifact. Class perceptions do vary within racial groups; a few white names read as below middle class and a few Black names above it. But the distributions barely overlap, so names cannot be matched on perceived class without leaving the racially distinctive pool altogether. More directly, their fifth study tells respondents the person's education explicitly. Told that a named person holds a high school degree, a state-school bachelor's degree or an Ivy League bachelor's degree, respondents still place Black-named individuals in a lower social class at every level. The gap narrows sharply at the Ivy League level and remains statistically significant (0.05 p-value).

The authors decline to rule on the question the critique turns on, and they are right to. If class inference is itself a mechanism of racial discrimination, netting it out would be post-treatment bias. Their bottom line is that measured discrimination probably reflects both racial and class bias rather than either alone.

Where this leaves the canonical number. Their headline advantage of 50 percent more callbacks is very likely too large as an estimate of race as such. Nobody has produced the decomposition that would say by how much, and neither perception study links per-name perception rates to per-name callback rates. Both also run on prompted survey respondents rather than on hiring managers reading a resume with nobody asking — a real external-validity gap, and one that runs in both directions.

The bias that runs the other way

Every criticism above says the audit number is too big. Bohren, Hull and Imas argue the opposite, and have not been answered.[11]

Their point is about what holding the resume fixed accomplishes. It identifies the causal effect of the race signal at *that one hiring decision*. By construction it excludes any disparity that arose because discrimination somewhere earlier altered what is on the resume. A conventional audit identifies the direct component and nets out everything systemic.

Their two-node field experiment makes this concrete. They worked in the automotive sector, the industry Kline, Rose and Walters had flagged as the most discriminatory. They sent 1,001 applications, split between resumes with no prior work experience and resumes with one line of it. Among applicants with no experience, 25.6 percent of white-named resumes drew a callback against a far lower Black-named rate, a gap of 12.4 percentage points. Adding one line of relevant experience shrank the gap to 4.2 percentage points, which is not statistically significant. Experience itself raised callbacks by 9.5 percentage points.

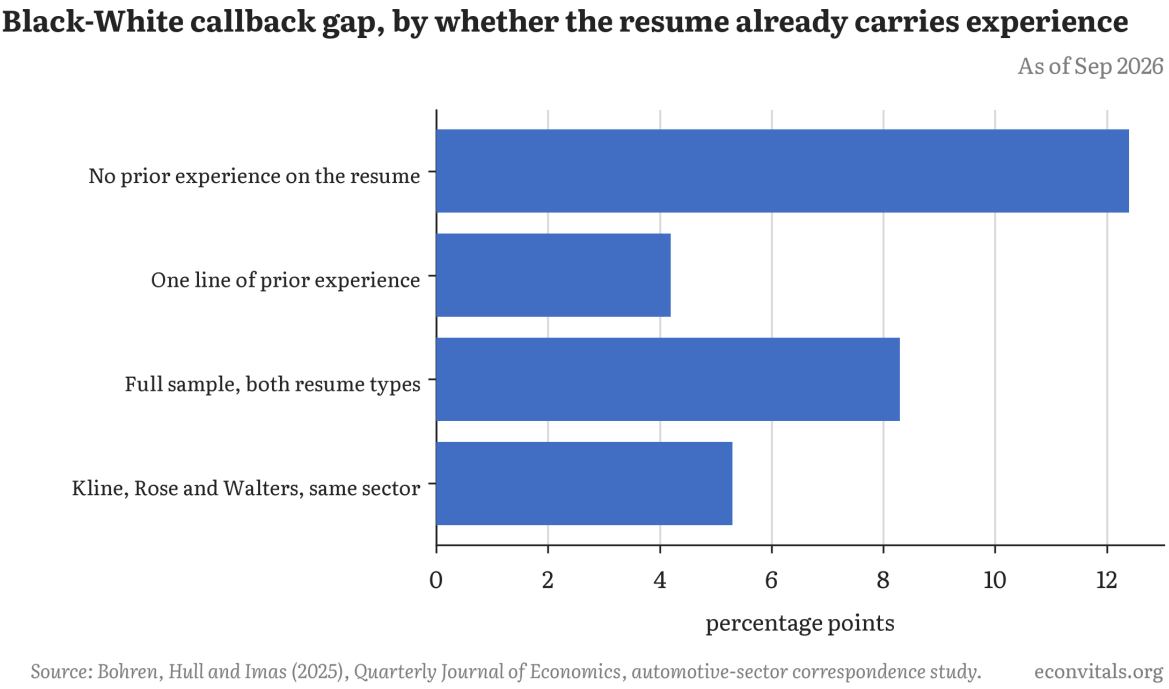


Figure 8. Holding the resume fixed at no experience the gap is 12.4 percentage points; adding one line of experience shrinks it to 4.2 percentage points, which is not significant. Because a conventional audit conditions on experience, reading only its estimate understates total discrimination by about 50 percent.
Source: Bohren, Hull and Imas (2025), Quarterly Journal of Economics, automotive-sector correspondence study.

Notice what that pattern does inside this sector. The inexperienced-resume gap is much larger than anything in the older literature. The experienced-resume gap is close to the 5.3 percentage points that Kline, Rose and Walters find for the same sector, where every resume carried experience. The experienced-resume gap is not statistically significant, so the contrast between the two rests on an imprecise estimate. Read narrowly, the pattern says that in one high-discrimination sector the gap is larger on resumes with no experience — the resumes the older studies did not send.

Because early discrimination denies Black applicants the experience that later employers reward, the first node feeds the second. At the median local market thickness the modeled second-node hiring gap is about 7.5 percentage points, and more than half of it is systemic rather than direct. Reading only the conventional direct-discrimination estimate would have understated total discrimination by about 50 percent — and by more if experience is shaped at

more than two hiring nodes.

Their gender experiment pushes the logic further. With recommendation letters held fixed, gender gaps in rated hireability and prospective wage are insignificant. Let the letters be generated from the applicant's own gendered materials and a wage gap of 21 percent appears. Total gender discrimination is 29 percent of a standard deviation in hiring likelihood, of which 27 percent of a standard deviation runs through the letters alone. Essentially all of it is systemic, and a design that holds materials fixed cannot see any of it.

The caveats matter. The second round is modeled, not observed. No applicant actually gained experience and reapplied, and the conversion rates come from a survey of 107 hiring managers rather than from realized hires. The systemic share swings from 11 percent in thin markets to considerably more at the median. And the firms were deliberately drawn from the sector already known to discriminate most.

Two biases, opposite signs, both real

Two of the economics papers here point in opposite directions. Fryer and Levitt's evidence implies the audit treatment is loaded with class, so the callback gap overstates the cost of race as such. Bohren, Hull and Imas imply the reverse: holding the application fixed removes the upstream channels through which race shaped it.

Both can hold at once, and most likely both do. **The audit number is too large as an estimate of the pure racial parameter and too small as an estimate of race's total contribution to hiring disparities.** Nobody has measured either bias well enough to say which dominates. Anyone who cites one without the other is telling half the story.

An average is not an employer

The third serious criticism is that a pooled callback gap describes nobody. It is correct, and the response has been to extend the method rather than abandon it.

Kline and Walters model callbacks as job-specific Bernoulli trials drawn from an unknown distribution and derive sharp bounds on the share of jobs that discriminate.[12] Re-analyzing Bertrand and Mullainathan's 1,112 job openings, at least 13 percent of jobs discriminate against Black applicants. In Nunley and coauthors' data the bound is at least 17 percent of jobs. A two-type mixed logit fitted independently lands close to the same figure (17 percent of jobs).

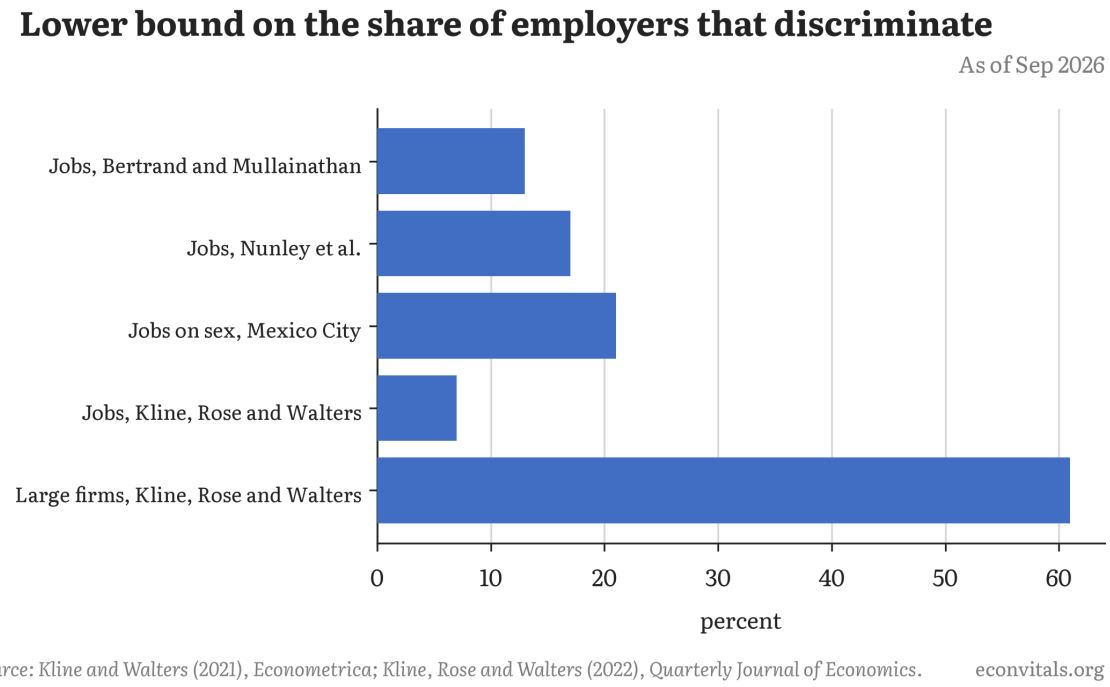


Figure 9. The average gap is produced by a minority of workplaces: at least 13 percent of jobs in the original experiment and 17 percent of jobs in Nunley et al., yet at least 61 percent of the 108 large firms in the biggest experiment discriminate somewhere in their hiring.

Source: Kline and Walters (2021), Econometrica; Kline, Rose and Walters (2022), Quarterly Journal of Economics.

Among the firms that do discriminate the behavior is not marginal. The estimated odds of a callback are roughly 53 times higher for a white than a Black application. The job-level standard deviation of the gap in Bertrand and Mullainathan is 0.082 callback-probability points, more than twice its mean. The higher moments show a long tail: most jobs barely discriminate and a few discriminate heavily.

This heterogeneity has forensic value. Of Bertrand and Mullainathan employers that called back two white applicants and neither Black one, the bound on discriminators is 72 percent of such jobs. In Nunley and coauthors the equivalent bound is 85 percent of such jobs, and a three-white design pushes it to 90 percent of such jobs. A regulator using the standard four-application design detects almost nothing, because a classical one-percent Fisher exact test on two-and-two corresponds to a posterior probability of discrimination of only 33 percent. Send ten applications per job and a regulator could detect roughly 7.5 percent of discriminating jobs while investigating fewer than 0.2 percent of the innocent.

The same result at scale

Kline, Rose and Walters recover the same structure with firm-level precision. After correcting for sampling error, the standard deviation of racial contact gaps across their 108 large employers is 1.9 percentage points — almost as large as the average gap itself. The spread across U.S. states is 0.8 percentage points, less than half of that.

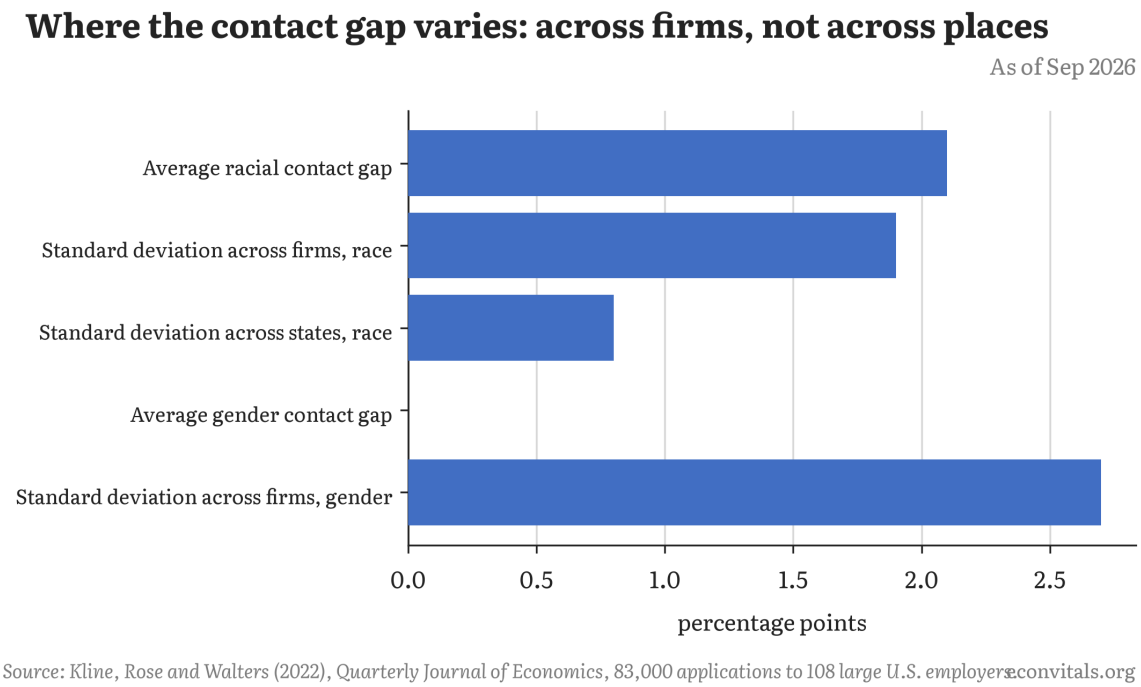


Figure 10. The between-firm standard deviation of racial contact gaps, 1.9 percentage points, is almost as large as the average gap itself, while the spread across states is 0.8 percentage points. The average gender gap is essentially zero and its between-firm spread is 2.7 percentage points.

Source: Kline, Rose and Walters (2022), Quarterly Journal of Economics, 83,000 applications to 108 large U.S. employers.

Discrimination is a property of firms, not places. Two-digit industry explains roughly 58 percent of the cross-firm variation. Firms in the top quintile account for 46 percent of all contacts lost to Black applicants. Empirical Bayes bounds imply at least 61 percent discriminate somewhere in their hiring, and controlling the false discovery rate identifies 23 companies individually. The most discriminatory sector is automotive dealers and service, at a posterior mean gap of 4.2 percentage points — which is why Bohren, Hull and Imas went there.

The gender result is the cleanest demonstration that an average can be actively misleading. The average male-female contact gap is essentially zero. The between-firm standard deviation of gender gaps is 2.7 percentage points, larger than the racial one. Some firms favor men, others favor women, and the average of the two is a number describing no employer in the sample.

The mechanism, seen directly

Agan and Starr's ban-the-box experiment supplies the sharpest evidence on what drives the gap, because it varies the employer's information rather than the applicant.[13] They sent 14,637 applications to 4,291 establishments in New Jersey and New York City. Pooled, white applicants received 23 percent more callbacks.

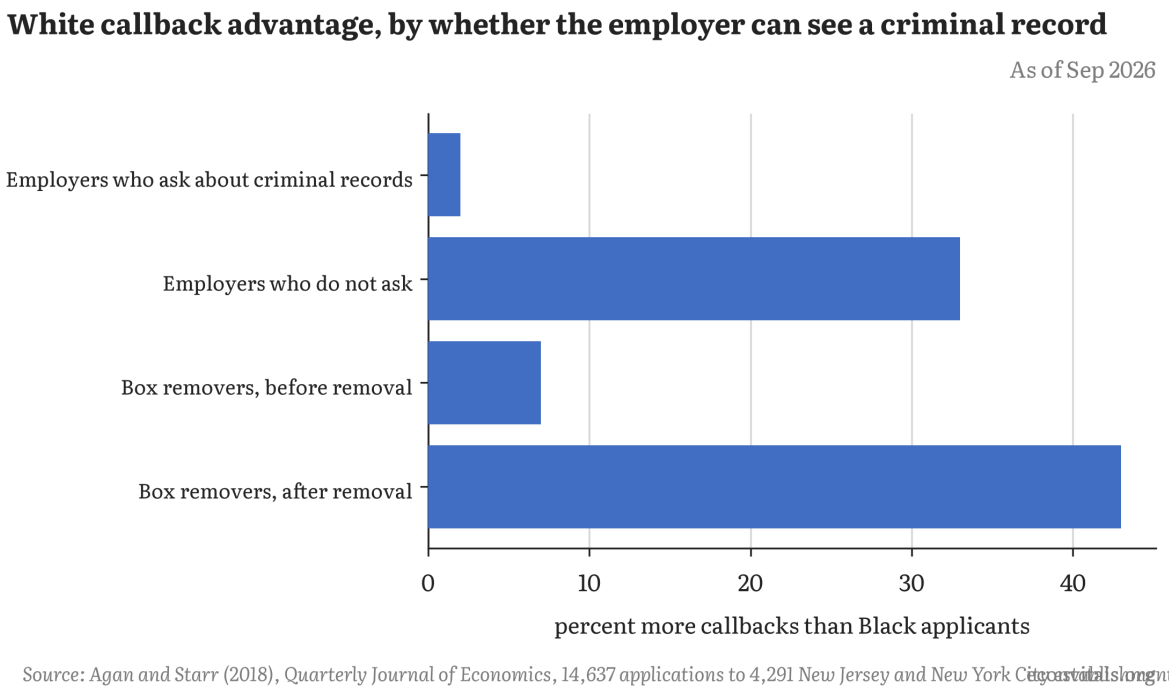


Figure 11. Where employers ask about records the race gap is 0.3 percentage points; where they do not, white applicants receive 33 percent. The same employers before and after dropping the question show the gap widening by 3.6 percentage points.

Source: Agan and Starr (2018), Quarterly Journal of Economics, 14,637 applications to 4,291 New Jersey and New York City establishments.

The pooled figure conceals two regimes. At employers whose applications asked about criminal records the race gap was 0.3 percentage points, economically and statistically insignificant. At employers who did not ask, white applicants received 33 percent more callbacks — an order of magnitude larger. Geography splits the same way: 4.5 percentage points in New Jersey against a small and insignificant gap in New York City.

The panel makes it causal rather than correlational. When box-remover firms lost the question, the race gap at those same firms widened by 3.6 percentage points, with a triple-difference estimate of 3.9 percentage points once trends at unaffected employers are differenced out.

What employers were doing is not accurate statistical inference. Backing priors out of the callback rates implies employers act as if the Black-white gap in felony convictions is 71 percentage points. The largest real-world gap the authors could find is about 32 percentage points. And employers do not respond at all to a GED, which predicts a criminal record far better than race does. Among GED holders, 35.8 percent had a conviction by early adulthood, against a much lower rate for diploma holders. This is the closest thing in the correspondence literature to a direct read on mechanism, and it points to inaccurate beliefs rather than to correct statistical discrimination or to pure animus. The in-person audits calibrate the same inference from the other side. A white applicant who disclosed a felony drug conviction drew a

positive response 17.2 percent of the time. That is no worse than an equally qualified Black applicant with a clean record.[14]

Agan and Starr also run the translation to labor markets, which almost no audit paper does. Applied to the real-world distribution of criminal records, their callback pattern implies a widening of the Black-white gap of 2.8 percentage points, a quarter of the study's overall callback rate of 11.7 percent. Using survey-based conviction rates instead raises it to 4.2 percentage points. They are candid that this is arithmetic on the experiment's own numbers, not a measured labor-market effect, and they cite Heckman and Siegelman when saying so.

From callbacks to jobs: the criticism still standing

Every correspondence design in this briefing ends at the callback. Bertrand and Mullainathan state that they cannot translate their results into gaps in hiring rates or earnings. Agan and Starr call the inability to observe ultimate hiring their most significant limitation. Kline, Rose and Walters observe only employer contact within thirty days. Quillian and coauthors say the evidence covers the point of hire only, not wage setting, promotion or termination.

Heckman's version is the strongest, and not merely about measurement. A well-designed audit could uncover many individual firms that discriminate while the marginal effect on the wages of employed workers is zero. Rejected applicants apply elsewhere, and minorities sort away from the worst firms. Lang and Kahn-Lang Spitzer make the same point from the other direction. The experimental evidence of discrimination in hiring is substantial, they write, and economics still knows very little about how much of the Black-white gap in wages and employment discrimination accounts for.[15] Their vocabulary is the useful part: taste-based discrimination, accurate statistical discrimination and statistical discrimination on inaccurate beliefs carry different policy implications, and a callback gap does not distinguish them. Kline, Rose and Walters decline to attempt that separation. Fryer and Levitt say they cannot make it. Among the papers read for this briefing, only Agan and Starr get close, and only by an indirect argument about employer priors.

One design in this briefing does reach the offer stage. Pager, Western and Bonikowski's New York City in-person audit sent matched teams of live testers to apply for 340 real entry-level jobs over nine months in 2004.[14] The testers were white, Latino and Black, carried equivalent resumes, and the study recorded job offers as well as callbacks. The white tester drew a positive response 31.0 percent of the time against 15.2 percent for the Black tester, a ratio of 2.04. That is the correspondence literature's result, reproduced where the correspondence literature cannot look. Two things stop it closing the question. Positive responses split roughly evenly between callbacks and on-the-spot offers, and the authors report offers distributed more evenly by race than callbacks. The gap sits at the screen, which is where the critics say it sits. And an in-person audit is the design Heckman and Siegelman wrote against: live testers can differ in ways no resume records, which is why the correspondence method displaced them. The offer-stage evidence is real and much older than the framing of this debate usually

allows. It is still not a firm's measured gap linked to that firm's own hires.

A circumstantial answer points the other way from the objection. Quillian and coauthors point to the aggregates as their meta-analysis found them. Racial gaps in unemployment have shown little change since 1980, and the Black-white gap in labor force participation among young men widened over that period. Both series moved as the flat experimental trend would predict. That is consistency rather than identification, and it describes the labor market as that meta-analysis found it rather than today's.

The firm-level test is now being attempted. Onorato, Stavreva, Sterling and Yimfor match the firms in the Kline, Rose and Walters experiment to employment records and ask whether a firm's measured callback gap predicts who actually works there.[16] Their reported answer is that it does. Firms with larger callback gaps employ fewer Black workers, and the implied employment gap is a multiple of what the callback gap alone would predict. The relationship is strongest among the firms that discriminated most in the experiment. Two readings are live and the paper cannot fully separate them — discrimination continuing past the resume screen, or Black applicants sorting away from firms with a reputation for it. Both are awkward for the claim that callback evidence is an artifact confined to the screening stage.

That paper could not be obtained for this briefing. It sits behind a repository that blocks automated retrieval, and nothing above comes from reading it — only from its research description. It is a working paper besides. Treat it as the shape of the answer that is coming, not the answer. The same applies, more mildly, to Lang and Kahn-Lang Spitzer, of which only the abstract could be retrieved; no passage of theirs is quoted.

Where this question actually stands. The callbacks-to-jobs link is the weakest joint in the chain and has been since 1998. The in-person audits reached the offer stage and found the same gap; what nobody has done is tie one employer's measured gap to whom that employer hires. Anyone who tells you the audit literature has established the size of labor-market discrimination is overclaiming. Anyone who calls a callback gap uninformative has to explain the firm-level link now being reported.

The scoreboard

None of the criticisms has retired the method. Each has been given a test of its own, and the tests have generally made the designs better rather than the finding smaller.

The main criticisms of audit studies and where each one stands

Criticism	Raised by	How it was tested	Where it stands
Unequal variances of unobservab...	Heckman and Siegelman (1993); ...	Neumark's heteroskedastic probi...	Real, and testable. On the origina...
Distinctively Black names signal ...	Fryer and Levitt (2004); Gaddis (2...	Birth-record name indices, name...	Substantially conceded. Class tra...
Callbacks are not hires, so the ga...	Heckman (1998); the studies' own ...	Employer surveys on callback-to-...	Open. No audit design observes h...
Audits sample newspaper and on...	Heckman (1998); Granovetter via ...	Comparison with survey evidenc...	Conceded as a limit on external v...
An average gap says nothing abo...	Kline and Walters (2021)	Job-level mixture models and em...	Answered by extending the meth...
Measuring one hiring node misse...	Bohren, Hull and Imas (2025)	A two-node design that randomiz...	Accepted, and it cuts against the ...

Figure 12. Each criticism has been given a test of its own; none of them has retired the method, and two of them imply the conventional estimate is too small rather than too large.

Source: Heckman (1998); Neumark (2012); Neumark and Rich (2019); Gaddis (2017); Crabtree et al. (2022); Kline and Walters (2021); Bohren, Hull and Imas (2025).

Two of the criticisms have been substantially conceded and both bear on magnitude rather than existence. The class content of names is real and unresolved. The average gap describes no employer: the top quintile of firms produces 46 percent of the shortfall. One — the identification critique — is real and testable. On the flagship data it turns the finding up rather than down, while genuinely overturning a majority of the older labor-market estimates. One is conceded as a limit on external validity and not on internal validity: audits sample advertised openings, which is a minority of how people find work. One cuts entirely against the critics, since measuring a single hiring node misses the discrimination that runs through the resume itself. And one is open: callbacks are not hires.

The finding that employers respond differently to identical applications carrying different racial signals is about as robust as an empirical result in economics gets. It has survived a change of medium, an order-of-magnitude increase in sample size, a change from newspaper ads to automated hiring portals, pre-registration, and a statistical correction designed by its critics. What it does not do, and never did, is tell you how much of the Black-white employment or wage gap is discrimination.

What to monitor

- **The Heckman-Siegelman correction applied to a modern within-firm design.** Kline, Rose and Walters randomize applicant quality inside each firm, exactly the condition Neumark's method requires, and nobody has run the heteroskedastic probit on it. That exercise would settle more of this debate than any new experiment.
- **Whether the callback-to-workforce link replicates and survives peer review.** The Onorato and coauthors result is the first firm-level link from an experimental callback gap to a realized workforce. The in-person audits reached the offer stage decades ago, but never tied a gap to one employer's own hires. Watch for the published version, for whether it can separate continuing discrimination from applicant sorting, and for whether anyone reproduces it on a different firm sample.

- **A meta-analysis with post-2015 fieldwork.** The flat U.S. trend rests on fieldwork that stops before the largest experiment in this literature. Whether the level gap has moved since is unknown rather than answered.
- **A design that crosses names with explicit socioeconomic signals.** The class confound is conceded and unquantified. An experiment varying perceived class independently of the name would produce the decomposition the perception studies cannot.
- **Whether anyone measures the second hiring node directly.** Bohren, Hull and Imas build their second round from surveyed conversion rates and scraped vacancy counts. Following real applicants forward would make the systemic-discrimination result a measured magnitude rather than a modeled one.
- **The regulatory design question.** Kline and Walters show a ten-application design can identify individual discriminators with a small false-positive rate. The technical obstacle is gone; two others are not. The EEOC opened nearly 600 systemic-discrimination investigations in 2020, many of them into practices such as word-of-mouth recruitment that a conventional correspondence or audit study would not detect. And whether an enforcement body adopts a detection design of this kind turns on enforcement priorities, which move with each administration, rather than on statistics.

Notes

1. 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. https://www.nber.org/papers/w9873
2. 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. https://www.pnas.org/doi/10.1073/pnas.1706255114
3. Kline, Patrick M., Evan K. Rose and Christopher R. Walters (2022). Systemic Discrimination Among Large U.S. Employers. Quarterly Journal of Economics 137(4): 1963-2036. https://www.nber.org/papers/w29053
4. Lippens, Louis, Siel Vermeiren and Stijn Baert (2023). The State of Hiring Discrimination: A Meta-Analysis of (Almost) All Recent Correspondence Experiments. European Economic Review 151: 104315. https://www.sciencedirect.com/science/article/pii/S0014292122001957
5. Heckman, James J. (1998). Detecting Discrimination. Journal of Economic Perspectives 12(2): 101-116. https://www.aeaweb.org/articles?id=10.1257/jep.12.2.101
6. Fryer, Roland G. Jr. and Steven D. Levitt (2004). The Causes and Consequences of Distinctively Black Names. Quarterly Journal of Economics 119(3): 767-805. https://www.nber.org/papers/w9938
7. Neumark, David (2012). Detecting Discrimination in Audit and Correspondence Studies. Journal of Human Resources 47(4): 1128-1157. https://jhr.uwpress.org/content/47/4/1128
8. Neumark, David and Judith Rich (2019). Do Field Experiments on Labor and Housing Markets Overstate Discrimination? A Re-examination of the Evidence. ILR Review 72(1): 223-252. https://journals.sagepub.com/doi/abs/10.1177/0019793918759665
9. Gaddis, S. Michael (2017). How Black Are Lakisha and Jamal? Racial Perceptions from Names Used in Correspondence Audit Studies. Sociological Science 4: 469-489. https://sociologicalscience.com/articles-v4-19-469/
10. Crabtree, Charles, S. Michael Gaddis, John B. Holbein and Edvard Nergård Larsen (2022). Racially Distinctive Names Signal Both Race/Ethnicity and Social Class. Sociological Science 9: 454-472.

https://sociologicalscience.com/articles-v9-18-454/

11. Bohren, J. Aislinn, Peter Hull and Alex Imas (2025). Systemic Discrimination: Theory and Measurement. Quarterly Journal of Economics 140(3): 1743-1799. https://academic.oup.com/qje/article/140/3/1743/8123617

12. Kline, Patrick and Christopher Walters (2021). Reasonable Doubt: Experimental Detection of Job-Level Employment Discrimination. Econometrica 89(2): 765-792. https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA17489

13. Agan, Amanda Y. and Sonja B. Starr (2018). Ban the Box, Criminal Records, and Racial Discrimination: A Field Experiment. Quarterly Journal of Economics 133(1): 191-235. https://academic.oup.com/qje/article-abstract/133/1/191/4060073

14. Pager, Devah, Bruce Western and Bart Bonikowski (2009). Discrimination in a Low-Wage Labor Market: A Field Experiment. American Sociological Review 74(5): 777-799; read in Pager and Western, Investigating Prisoner Reentry, National Institute of Justice final report. https://www.ojp.gov/pdffiles1/nij/grants/228584.pdf

15. Lang, Kevin and Ariella Kahn-Lang Spitzer (2020). Race Discrimination: An Economic Perspective. Journal of Economic Perspectives 34(2): 68-89. https://www.aeaweb.org/articles?id=10.1257/jep.34.2.68

16. Onorato, Donato, Kamelia Stavreva, Adina Sterling and Emmanuel Yimfor (2025). From Callbacks to Careers: Do Audit Studies Predict Workforce Diversity? Columbia Business School Research Paper No. 6022958. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6022958

Principal sources

Principal sources behind this briefing, drawn from the evidence it cites. Public sources carry their publication and link; desk notes are named by source and date.

- How Black Are Lakisha and Jamal? Racial Perceptions from Names Used in Correspondence Audit Studies, S. Michael Gaddis, Sociological Science 4: 469-489, 2017-09-06. https://sociologicalscience.com/articles-v4-19-469/
- Racially Distinctive Names Signal Both Race/Ethnicity and Social Class, Charles Crabtree, S. Michael Gaddis, John B. Holbein and Edvard Nergard Larsen, Sociological Science 9: 454-472, 2022-12-12. https://sociologicalscience.com/articles-v9-18-454/
- The Causes and Consequences of Distinctively Black Names, Roland G. Fryer, Jr. and Steven D. Levitt, Quarterly Journal of Economics 119(3): 767-805, 2004-08-01. https://www.nber.org/papers/w9938
- Systemic Discrimination: Theory and Measurement, J. Aislinn Bohren, Peter Hull and Alex Imas, Quarterly Journal of Economics 140(3): 1743-1799 (NBER Working Paper 29820, March 2022, revised December 2024), 2025-08-01. https://academic.oup.com/qje/article/140/3/1743/8123617
- Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination, Marianne Bertrand and Sendhil Mullainathan, American Economic Review 94(4): 991-1013, 2004. https://www.nber.org/papers/w9873
- Meta-analysis of field experiments shows no change in racial discrimination in hiring over time, Lincoln Quillian, Devah Pager, Ole Hexel and Arnfinn H. Midtboen, Proceedings of the National Academy of Sciences 114(41): 10870-10875, 2017. https://www.pnas.org/doi/10.1073/pnas.1706255114
- Systemic Discrimination Among Large U.S. Employers, Patrick M. Kline, Evan K. Rose and Christopher R. Walters, Quarterly Journal of Economics 137(4): 1963-2036 (NBER Working Paper 29053), 2022. https://www.nber.org/papers/w29053
- The State of Hiring Discrimination: A Meta-Analysis of (Almost) All Recent Correspondence Experiments, Louis Lippens, Siel Vermeiren and Stijn Baert, European Economic Review 151: 104315, 2023. https://www.sciencedirect.com/science/article/pii/S0014292122001957
- Reasonable Doubt: Experimental Detection of Job-Level Employment Discrimination, Patrick Kline and Christopher Walters, Econometrica 89(2): 765-792 (NBER Working Paper 26861), 2021-03. https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA17489
- Ban the Box, Criminal Records, and Racial Discrimination: A Field Experiment, Amanda Y. Agan and Sonja B. Starr, Quarterly Journal of Economics 133(1): 191-235, 2018-02. https://academic.oup.com/qje/article-abstract/133/1/191/4060073
- Detecting Discrimination, James J. Heckman, Journal of Economic Perspectives 12(2): 101-116, 1998. https://www.aeaweb.org/articles?id=10.1257/jep.12.2.101
- Detecting Discrimination in Audit and Correspondence Studies, David Neumark, Journal of Human Resources 47(4): 1128-1157 (NBER Working Paper 16448), 2012. https://jhr.uwpress.org/content/47/4/1128
- Do Field Experiments on Labor and Housing Markets Overstate Discrimination? A Re-examination of the Evidence, David Neumark and Judith Rich, ILR Review 72(1): 223-252 (NBER Working Paper 22278), 2019. https://journals.sagepub.com/doi/abs/10.1177/0019793918759665
- Investigating Prisoner Reentry: The Impact of Conviction Status on the Employment Prospects of Young Men, Devah Pager and Bruce Western, National Institute of Justice final report 228584 (the New York City in-person audit published as Pager, Western and Bonikowski, 'Discrimination in a Low-Wage Labor Market: A Field Experiment,' American Sociological Review 74(5): 777-799), 2009. https://www.ojp.gov/pdffiles1/nij/grants/228584.pdf

Bound sources

The documents read for “What The Audit Studies Found, And What The Critics Got Right”, bound in full after the memo. Each begins with a cover page; page numbers are the binder’s. by Claude, may contain mistakes.

Section 1 · What the Audit Studies Actually Found

E-lit-005: Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination 28

E-lit-006: Meta-analysis of field experiments shows no change in racial discrimination in hiring over time 52

E-lit-007: Systemic Discrimination Among Large U.S. Employers 59

E-lit-008: The State of Hiring Discrimination: A Meta-Analysis of (Almost) All Recent Correspondence Experiments 176

Section 2 · The Identification Critique: Heckman-Siegelman, Neumark's Correction, and What Survives It

E-lit-013: Detecting Discrimination 204

E-lit-014: Detecting Discrimination in Audit and Correspondence Studies 222

E-lit-015: Do Field Experiments on Labor and Housing Markets Overstate Discrimination? A Re-examination of the Evidence 271

E-lit-007: Systemic Discrimination Among Large U.S. Employers 311

Section 3 · What Is Actually Being Manipulated: Race, Class, and the Resume Held Fixed

E-lit-001: How Black Are Lakisha and Jamal? Racial Perceptions from Names Used in Correspondence Audit Studies 428

E-lit-002: Racially Distinctive Names Signal Both Race/Ethnicity and Social Class 450

E-lit-003: The Causes and Consequences of Distinctively Black Names 470

E-lit-004: Systemic Discrimination: Theory and Measurement 511

Section 4 · From Callbacks to Labor Markets: Which Employers, How Much, and Why

E-lit-009: Reasonable Doubt: Experimental Detection of Job-Level Employment Discrimination 578

E-lit-010: From Callbacks to Careers: Do Audit Studies Predict Workforce Diversity? 625

E-lit-011: Ban the Box, Criminal Records, and Racial Discrimination: A Field Experiment 627

E-lit-012: Race Discrimination: An Economic Perspective 674

E-lit-016: Investigating Prisoner Reentry: The Impact of Conviction Status on the Employment Prospects of Young Men 678

E-lit-005: Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination

Marianne Bertrand and Sendhil Mullainathan | American Economic Review 94(4): 991-1013 | 2004

https://www.nber.org/papers/w9873

About this source

Bertrand and Mullainathan find a large, precisely estimated callback penalty attached to distinctively Black names. Resumes with white names were called back 9.65 percent of the time and otherwise identical resumes with Black names 6.45 percent of the time: a gap of 3.20 percentage points, or 50 percent in relative terms, which they gloss as a white applicant needing about 10 resumes per callback against about 15 for a Black applicant. The gap is essentially the same in Chicago (49 percent) and Boston (50 percent), for men (52 percent) and women (49 percent), and they cannot reject that it is the same across every occupation and industry in the study. The second finding matters as much as the first: improving a resume raised the white callback rate by 2.29 points (27 percent, significant) but the Black rate by only 0.51 points (8 percent, insignificant), so the absolute gap widens with credentials; under an objective predicted-quality ranking Black applicants do gain, but by less (ratio 1.60 versus 1.89). Employers who advertise as Equal Opportunity Employers and federal contractors discriminate no less. Living in a whiter, richer or more educated zip code raises callbacks for everyone but helps Black-named applicants no more than white-named ones. The authors test the leading benign explanation — that the names proxy for social class — using mothers' education from Massachusetts birth certificates, and find the rank correlation between a name's callback rate and its mothers' education runs the wrong way in all four race-sex groups. For the section's hypothesis this paper is the anchor point: it is the source of the canonical 50 percent figure, and its 3.2-point absolute gap and 9.65/6.45 base rates are the benchmark that later, better-powered experiments are measured against.

Bound as the original PDF. Retrieved 2026-09-10. sha256 e2c6c2854c9d497f...

Bound into an Economic Briefing · by Claude, may contain mistakes

Full document begins on the next page.

E-lit-005: https://www.nber.org/papers/w9873

Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination

By MARIANNE BERTRAND AND SENDHIL MULLAINATHAN*

We study race in the labor market by sending fictitious resumes to help-wanted ads in Boston and Chicago newspapers. To manipulate perceived race, resumes are randomly assigned African-American- or White-sounding names. White names receive 50 percent more callbacks for interviews. Callbacks are also more responsive to resume quality for White names than for African-American ones. The racial gap is uniform across occupation, industry, and employer size. We also find little evidence that employers are inferring social class from the names. Differential treatment by race still appears to still be prominent in the U.S. labor market. (*JEL* J71, J64).

Every measure of economic success reveals significant racial inequality in the U.S. labor market. Compared to Whites, African-Americans are twice as likely to be unemployed and earn nearly 25 percent less when they are employed (Council of Economic Advisers, 1998). This inequality has sparked a debate as to whether employers treat members of different races differentially. When faced with observably similar African-American and White applicants, do they favor the White one? Some argue yes, citing either employer prejudice or employer perception that race signals lower productivity. Others argue that differential treatment by race is a relic of the past, eliminated by some combination of employer enlightenment, affirmative action programs and the profit-maximization motive. In fact, many in this latter camp even feel that stringent enforcement of affirmative action programs has produced an environment of reverse discrimination. They would argue that faced with identical candidates, employers might favor the African-American one.[1] Data limitations make it difficult to empirically test these views. Since researchers possess far less data than employers do, White and African-American workers that appear similar to researchers may look very different to employers. So any racial difference in labor market outcomes could just as easily be attributed to differences that are observable to employers but unobservable to researchers.

To circumvent this difficulty, we conduct a field experiment that builds on the correspondence testing methodology that has been primarily used in the past to study minority outcomes in the United Kingdom.[2] We send resumes in response to help-wanted ads in Chicago and Boston newspapers and measure callback for interview for each sent resume. We

* Bertrand: Graduate School of Business, University of Chicago, 1101 E. 58th Street, R0 229D, Chicago, IL 60637, NBER, and CEPR (e-mail: marianne.bertrand@gsb.uchicago.edu); Mullainathan: Department of Economics, Massachusetts Institute of Technology, 50 Memorial Drive, E52-380a, Cambridge, MA 02142, and NBER (e-mail: mullain@mit.edu). David Abrams, Victoria Bede, Simone Berkowitz, Hong Chung, Almudena Fernandez, Mary Anne Guediguian, Christine Jaw, Richa Maheswari, Beverley Martis, Alison Tisza, Grant Whitehorn, and Christine Yee provided excellent research assistance. We are also grateful to numerous colleagues and seminar participants for very helpful comments.

[1] This camp often explains the poor performance of African-Americans in terms of supply factors. If African-Americans lack many basic skills entering the labor market, then they will perform worse, even with parity or favoritism in hiring.

[2] See Roger Jowell and Patricia Prescott-Clarke (1970), Jim Hubbuck and Simon Carter (1980), Colin Brown and Pat Gay (1985), and Peter A. Riach and Judith Rich (1991). One caveat is that some of these studies fail to fully match skills between minority and nonminority resumes. For example some impose differential education background by racial origin. Doris Weichselbaumer (2003, 2004) studies the impact of sex-stereotypes and sexual orientation. Richard E. Nisbett and Dov Cohen (1996) perform a related field experiment to study how employers' response to a criminal past varies between the North and the South in the United States.

experimentally manipulate perception of race via the name of the fictitious job applicant. We randomly assign very White-sounding names (such as Emily Walsh or Greg Baker) to half the resumes and very African-American-sounding names (such as Lakisha Washington or Jamal Jones) to the other half. Because we are also interested in how credentials affect the racial gap in callback, we experimentally vary the quality of the resumes used in response to a given ad. Higher-quality applicants have on average a little more labor market experience and fewer holes in their employment history; they are also more likely to have an e-mail address, have completed some certification degree, possess foreign language skills, or have been awarded some honors.[3] In practice, we typically send four resumes in response to each ad: two higher-quality and two lower-quality ones. We randomly assign to one of the higher- and one of the lower-quality resumes an African-American-sounding name. In total, we respond to over 1,300 employment ads in the sales, administrative support, clerical, and customer services job categories and send nearly 5,000 resumes. The ads we respond to cover a large spectrum of job quality, from cashier work at retail establishments and clerical work in a mail room, to office and sales management positions.

We find large racial differences in callback rates.[4] Applicants with White names need to send about 10 resumes to get one callback whereas applicants with African-American names need to send about 15 resumes. This 50-percent gap in callback is statistically significant. A White name yields as many more callbacks as an additional eight years of experience on a resume. Since applicants' names are randomly assigned, this gap can only be attributed to the name manipulation.

Race also affects the reward to having a better resume. Whites with higher-quality resumes receive nearly 30-percent more callbacks than Whites with lower-quality resumes. On the other hand, having a higher-quality resume has a smaller effect for African-Americans. In other words, the gap between Whites and African-Americans widens with resume quality. While one may have expected improved credentials to alleviate employers' fear that African-American applicants are deficient in some unobservable skills, this is not the case in our data.[5]

The experiment also reveals several other aspects of the differential treatment by race. First, since we randomly assign applicants' postal addresses to the resumes, we can study the effect of neighborhood of residence on the likelihood of callback. We find that living in a wealthier (or more educated or Whiter) neighborhood increases callback rates. But, interestingly, African-Americans are not helped more than Whites by living in a "better" neighborhood. Second, the racial gap we measure in different industries does not appear correlated to Census-based measures of the racial gap in wages. The same is true for the racial gap we measure in different occupations. In fact, we find that the racial gaps in callback are statistically indistinguishable across all the occupation and industry categories covered in the experiment. Federal contractors, who are thought to be more severely constrained by affirmative action laws, do not treat the African-American resumes more preferentially; neither do larger employers or employers who explicitly state that they are "Equal Opportunity Employers." In Chicago, we find a slightly smaller racial gap when employers are located in more African-American neighborhoods.

The rest of the paper is organized as follows. Section I compares this experiment to earlier work on racial discrimination, and most notably to the labor market audit studies. We describe the experimental design in Section II and present the results in Section III, subsection A. In Section IV, we discuss possible interpretations of our results, focusing especially on two issues. First, we examine whether the

[3] In creating the higher-quality resumes, we deliberately make small changes in credentials so as to minimize the risk of overqualification.

[4] For ease of exposition, we refer to the effects uncovered in this experiment as racial differences. Technically, however, these effects are about the racial soundingness of names. We briefly discuss below the potential confounds between name and race. A more extensive discussion is offered in Section IV, subsection B.

[5] These results contrast with the view, mostly based on nonexperimental evidence, that African-Americans receive higher returns to skills. For example, estimating earnings regressions on several decades of Census data, James J. Heckman et al. (2001) show that African-Americans experience higher returns to a high school degree than Whites do.

race-specific names we have chosen might also proxy for social class above and beyond the race of the applicant. Using birth certificate data on mother's education for the different first names used in our sample, we find little relationship between social background and the name-specific callback rates.[6] Second, we discuss how our results map back to the different models of discrimination proposed in the economics literature. In doing so, we focus on two important results: the lower returns to credentials for African-Americans and the relative homogeneity of the racial gap across occupations and industries. We conclude that existing models do a poor job of explaining the full set of findings. Section V concludes.

I. Previous Research

With conventional labor force and household surveys, it is difficult to study whether differential treatment occurs in the labor market.[7] Armed only with survey data, researchers usually measure differential treatment by comparing the labor market performance of Whites and African-Americans (or men and women) for which they observe similar sets of skills. But such comparisons can be quite misleading. Standard labor force surveys do not contain all the characteristics that employers observe when hiring, promoting, or setting wages. So one can never be sure that the minority and nonminority workers being compared are truly similar from the employers' perspective. As a consequence, any measured differences in outcomes could be attributed to these unobserved (to the researcher) factors.

This difficulty with conventional data has led some authors to instead rely on pseudo-experiments.[8] Claudia Goldin and Cecilia Rouse (2000), for example, examine the effect of blind auditioning on the hiring process of orchestras. By observing the treatment of female candidates before and after the introduction of blind auditions, they try to measure the amount of sex discrimination. When such pseudo-experiments can be found, the resulting study can be very informative; but finding such experiments has proven to be extremely challenging.

A different set of studies, known as audit studies, attempts to place comparable minority and White actors into actual social and economic settings and measure how each group fares in these settings.[9] Labor market audit studies send comparable minority (African-American or Hispanic) and White auditors in for interviews and measure whether one is more likely to get the job than the other.[10] While the results vary somewhat across studies, minority auditors tend to perform worse on average: they are less likely to get called back for a second interview and, conditional on getting called back, less likely to get hired.

These audit studies provide some of the cleanest nonlaboratory evidence of differential treatment by race. But they also have weaknesses, most of which have been highlighted in Heckman and Siegelman (1992) and Heckman (1998). First, these studies require that both members of the auditor pair are identical in all dimensions that might affect productivity in employers' eyes, except for race. To accomplish this, researchers typically match auditors on several characteristics (height, weight, age, dialect, dressing style, hairdo) and train them for several days to coordinate interviewing styles. Yet, critics note that this is unlikely to erase the numerous differences that exist between the auditors in a pair.

Another weakness of the audit studies is that they are not double-blind. Auditors know the purpose of the study. As Turner et al. (1991)

[6] We also argue that a social class interpretation would find it hard to explain some of our findings, such as why living in a better neighborhood does not increase callback rates more for African-American names than for White names.

[7] See Joseph G. Altonji and Rebecca M. Blank (1999) for a detailed review of the existing literature on racial discrimination in the labor market.

[8] William A. Darity, Jr. and Patrick L. Mason (1998) describe an interesting nonexperimental study. Prior to the Civil Rights Act of 1964, employment ads would explicitly state racial biases, providing a direct measure of differential treatment. Of course, as Arrow (1998) mentions, discrimination was at that time "a fact too evident for detection."

[9] Michael Fix and Marjery A. Turner (1998) provide a survey of many such audit studies.

[10] Earlier hiring audit studies include Jerry M. Newman (1978) and Shelby J. McIntyre et al. (1980). Three more recent studies are Harry Cross et al. (1990), Franklin James and Steve W. DelCastillo (1991), and Turner et al. (1991). Heckman and Peter Siegelman (1992), Heckman (1998), and Altonji and Blank (1999) summarize these studies. See also David Neumark (1996) for a labor market audit study on gender discrimination.

note: "The first day of training also included an introduction to employment discrimination, equal employment opportunity, and a review of project design and methodology." This may generate conscious or subconscious motives among auditors to generate data consistent or inconsistent with their beliefs about race issues in America. As psychologists know very well, these demand effects can be quite strong. It is very difficult to insure that auditors will not want to do "a good job." Since they know the goal of the experiment, they can alter their behavior in front of employers to express (indirectly) their own views. Even a small belief by auditors that employers treat minorities differently can result in measured differences in treatment. This effect is further magnified by the fact that auditors are not in fact seeking jobs and are therefore more free to let their beliefs affect the interview process.

Finally, audit studies are extremely expensive, making it difficult to generate large enough samples to understand nuances and possible mitigating factors. Also, these budgetary constraints worsen the problem of mismatched auditor pairs. Cost considerations force the use of a limited number of pairs of auditors, meaning that any one mismatched pair can easily drive the results. In fact, these studies generally tend to find significant differences in outcomes across pairs.

Our study circumvents these problems. First, because we only rely on resumes and not people, we can be sure to generate comparability across race. In fact, since race is randomly assigned to each resume, the same resume will sometimes be associated with an African-American name and sometimes with a White name. This guarantees that any differences we find are caused solely by the race manipulation. Second, the use of paper resumes insulates us from demand effects. While the research assistants know the purpose of the study, our protocol allows little room for conscious or subconscious deviations from the set procedures. Moreover, we can objectively measure whether the randomization occurred as expected. This kind of objective measurement is impossible in the case of the previous audit studies. Finally, because of relatively low marginal cost, we can send out a large number of resumes. Besides giving us more precise estimates, this larger sample size also allows us to examine the nature of the differential treatment from many more angles.

II. Experimental Design

A. *Creating a Bank of Resumes*

The first step of the experimental design is to generate templates for the resumes to be sent. The challenge is to produce a set of realistic and representative resumes without using resumes that belong to actual job seekers. To achieve this goal, we start with resumes of actual job searchers but alter them sufficiently to create distinct resumes. The alterations maintain the structure and realism of the initial resumes without compromising their owners.

We begin with resumes posted on two job search Web sites as the basis for our artificial resumes.[11] While the resumes posted on these Web sites may not be completely representative of the average job seeker, they provide a practical approximation.[12] We restrict ourselves to people seeking employment in our experimental cities (Boston and Chicago). We also restrict ourselves to four occupational categories: sales, administrative support, clerical services, and customer services. Finally, we further restrict ourselves to resumes posted more than six months prior to the start of the experiment. We purge the selected resumes of the person's name and contact information.

During this process, we classify the resumes within each detailed occupational category into two groups: high and low quality. In judging resume quality, we use criteria such as labor market experience, career profile, existence of gaps in employment, and skills listed. Such a classification is admittedly subjective but it is made independently of any race assignment on the resumes (which occurs later in the experimental design). To further reinforce the quality gap between the two sets of resumes, we add to each high-quality resume a subset of the following features: summer or while-at-school employment experience, volunteering experience, extra computer skills, certification degrees, foreign language skills, honors, or some military

[11] The sites are www.careerbuilder.com and www.americasjobbank.com.

[12] In practice, we found large variation in skill levels among people posting their resumes on these sites.

experience. This resume quality manipulation needs to be somewhat subtle to avoid making a higher-quality job applicant overqualified for a given job. We try to avoid this problem by making sure that the features listed above are not all added at once to a given resume. This leaves us with a high-quality and a low-quality pool of resumes.[13]

To minimize similarity to actual job seekers, we use resumes from Boston job seekers to form templates for the resumes to be sent out in Chicago and use resumes from Chicago job seekers to form templates for the resumes to be sent out in Boston. To implement this migration, we alter the names of the schools and previous employers on the resumes. More specifically, for each Boston resume, we use the Chicago resumes to replace a Boston school with a Chicago school.[14] We also use the Chicago resumes to replace a Boston employer with a Chicago employer in the same industry. We use a similar procedure to migrate Chicago resumes to Boston.[15] This produces distinct but realistic looking resumes, similar in their education and career profiles to this subpopulation of job searchers.[16]

B. *Identities of Fictitious Applicants*

The next step is to generate identities for the fictitious job applicants: names, telephone numbers, postal addresses, and (possibly) e-mail addresses. The choice of names is crucial to our experiment.[17] To decide on which names are uniquely African-American and which are uniquely White, we use name frequency data calculated from birth certificates of all babies born in Massachusetts between 1974 and 1979. We tabulate these data by race to determine which names are distinctively White and which are distinctively African-American. Distinctive names are those that have the highest ratio of frequency in one racial group to frequency in the other racial group.

As a check of distinctiveness, we conducted a survey in various public areas in Chicago. Each respondent was asked to assess features of a person with a particular name, one of which is race. For each name, 30 respondents were asked to identify the name as either "White," "African-American," "Other," or "Cannot Tell." In general, the names led respondents to readily attribute the expected race for the person but there were a few exceptions and these names were disregarded.[18]

The final list of first names used for this study is shown in Appendix Table A1. The table reports the relative likelihood of the names for the Whites and African-Americans in the Massachusetts birth certificates data as well as the recognition rate in the field survey.[19] As Appendix Table A1 indicates, the African-American first names used in the experiment are quite common in the population. This suggests that by using these names as an indicator of race, we are actually covering a rather large segment of the African-American population.[20]

Applicants in each race/sex/city/resume quality cell are allocated the same phone number. This guarantees that we can precisely track employer callbacks in each of these cells. The phone lines we use are virtual ones with only a voice mailbox attached to them. A similar outgoing message is recorded on each of the voice mailboxes but each message is recorded by someone of the appropriate race and gender.

[13] In Section III, subsection B, and Table 3, we provide a detailed summary of resume characteristics by quality level.

[14] We try as much as possible to match high schools and colleges on quality and demographic characteristics.

[15] Note that for applicants with schooling or work experience outside of the Boston or Chicago areas, we leave the school or employer name unchanged.

[16] We also generate a set of different fonts, layouts, and cover letters to further differentiate the resumes. These are applied at the time the resumes are sent out.

[17] We chose name over other potential manipulations of race, such as affiliation with a minority group, because we felt such affiliations may especially convey more than race.

[18] For example, Maurice and Jerome are distinctively African-American names in a frequency sense yet are not perceived as such by many people.

[19] So many of names show a likelihood ratio of ∞ because there is censoring of the data at five births. If there are fewer than five babies in any race/name cell, it is censored (and we do not know whether a cell has zero or was censored). This is primarily a problem for the computation of how many African-American babies have "White" names.

[20] We also tried to use more White-sounding last names for White applicants and more African-American-sounding last names for African-American applicants. The last names used for White applicants are: Baker, Kelly, McCarthy, Murphy, Murray, O'Brien, Ryan, Sullivan, and Walsh. The last names used for African-American applicants are: Jackson, Jones, Robinson, Washington, and Williams.

Since we allocate the same phone number for applicants with different names, we cannot use a person name in the outgoing message.

While we do not expect positive feedback from an employer to take place via postal mail, resumes still need postal addresses. We therefore construct fictitious addresses based on real streets in Boston and Chicago using the White Pages. We select up to three addresses in each 5-digit zip code in Boston and Chicago. Within cities, we randomly assign addresses across all resumes. We also create eight e-mail addresses, four for Chicago and four for Boston.[21] These e-mail addresses are neutral with respect to both race and sex. Not all applicants are given an e-mail address. The e-mail addresses are used almost exclusively for the higher-quality resumes. This procedure leaves us with a bank of names, phone numbers, addresses, and e-mail addresses that we can assign to the template resumes when responding to the employment ads.

C. *Responding to Ads*

The experiment was carried out between July 2001 and January 2002 in Boston and between July 2001 and May 2002 in Chicago.[22] Over that period, we surveyed all employment ads in the Sunday editions of *The Boston Globe* and *The Chicago Tribune* in the sales, administrative support, and clerical and customer services sections. We eliminate any ad where applicants were asked to call or appear in person. In fact, most of the ads we surveyed in these job categories ask for applicants to fax in or (more rarely) mail in their resume. We log the name (when available) and contact information for each employer, along with any information on the position advertised and specific requirements (such as education, experience, or computer skills). We also record whether or not the ad explicitly states that the employer is an equal opportunity employer.

For each ad, we use the bank of resumes to sample four resumes (two high-quality and two low-quality) that fit the job description and requirements as closely as possible.[23] In some cases, we slightly alter the resumes to improve the quality of the match, such as by adding the knowledge of a specific software program.

One of the high- and one of the low-quality resumes selected are then drawn at random to receive African-American names, the other high- and low-quality resumes receive White names.[24] We use male and female names for sales jobs, whereas we use nearly exclusively female names for administrative and clerical jobs to increase callback rates.[25] Based on sex, race, city, and resume quality, we assign a resume the appropriate phone number. We also select at random a postal address. Finally, e-mail addresses are added to most of the high-quality resumes.[26] The final resumes are formatted, with fonts, layout, and cover letter style chosen at random. The resumes are then faxed (or in a few cases mailed) to the employer. All in all, we respond to more than 1,300 employment ads over the entire sample period and send close to 5,000 resumes.

D. *Measuring Responses*

We measure whether a given resume elicits a callback or e-mail back for an interview. For each phone or e-mail response, we use the content of the message left by the employer (name of the applicant, company name, telephone number for contact) to match the response to the corresponding resume-ad pair.[27] Any attempt by employers to contact applicants via postal mail cannot be measured in our experiment since the addresses are fictitious. Several human resource managers confirmed to us that

[21] The e-mail addresses are registered on Yahoo.com, Angelfire.com, or Hotmail.com.

[22] This period spans tighter and slacker labor markets. In our data, this is apparent as callback rates (and number of new ads) dropped after September 11, 2001. Interestingly, however, the racial gap we measure is the same across these two periods.

[23] In some instances, our resume bank does not have four resumes that are appropriate matches for a given ad. In such instances, we send only two resumes.

[24] Though the same names are repeatedly used in our experiment, we guarantee that no given ad receives multiple resumes with the same name.

[25] Male names were used for a few administrative jobs in the first month of the experiment.

[26] In the first month of the experiment, a few high-quality resumes were sent without e-mail addresses and a few low-quality resumes were given e-mail addresses. See Table 3 for details.

[27] Very few employers used e-mail to contact an applicant back.

TABLE 1—MEAN CALLBACK RATES BY RACIAL SOUNDINGNESS OF NAMES

	Percent callback for White names	Percent callback for African-American names	Ratio	Percent difference (*p*-value)
Sample:				
All sent resumes	9.65	6.45	1.50	3.20
	[2,435]	[2,435]		(0.0000)
Chicago	8.06	5.40	1.49	2.66
	[1,352]	[1,352]		(0.0057)
Boston	11.63	7.76	1.50	4.05
	[1,083]	[1,083]		(0.0023)
Females	9.89	6.63	1.49	3.26
	[1,860]	[1,886]		(0.0003)
Females in administrative jobs	10.46	6.55	1.60	3.91
	[1,358]	[1,359]		(0.0003)
Females in sales jobs	8.37	6.83	1.22	1.54
	[502]	[527]		(0.3523)
Males	8.87	5.83	1.52	3.04
	[575]	[549]		(0.0513)

Notes: The table reports, for the entire sample and different subsamples of sent resumes, the callback rates for applicants with a White-sounding name (column 1) an an African-American-sounding name (column 2), as well as the ratio (column 3) and difference (column 4) of these callback rates. In brackets in each cell is the number of resumes sent in that cell. Column 4 also reports the *p*-value for a test of proportion testing the null hypothesis that the callback rates are equal across racial groups.

employers rarely, if ever, contact applicants via postal mail to set up interviews.

E. *Weaknesses of the Experiment*

We have already highlighted the strengths of this experiment relative to previous audit studies. We now discuss its weaknesses. First, our outcome measure is crude, even relative to the previous audit studies. Ultimately, one cares about whether an applicant gets the job and about the wage offered conditional on getting the job. Our procedure, however, simply measures callbacks for interviews. To the extent that the search process has even moderate frictions, one would expect that reduced interview rates would translate into reduced job offers. However, we are not able to translate our results into gaps in hiring rates or gaps in earnings.

Another weakness is that the resumes do not directly report race but instead suggest race through personal names. This leads to various sources of concern. First, while the names are chosen to make race salient, some employers may simply not notice the names or not recognize their racial content. On a related note, because we are not assigning race but only race-specific names, our results are not representative of the average African-American (who may not have such a racially distinct name).[28] We return to this issue in Section IV, subsection B.

Finally, and this is an issue pervasive in both our study and the pair-matching audit studies, newspaper ads represent only one channel for job search. As is well known from previous work, social networks are another common means through which people find jobs and one that clearly cannot be studied here. This omission could qualitatively affect our results if African-Americans use social networks more or if employers who rely more on networks differentiate less by race.[29]

III. Results

A. *Is There a Racial Gap in Callback?*

Table 1 tabulates average callback rates by racial soundingness of names. Included in brackets under each rate is the number of resumes sent in that cell. Row 1 presents our results for the full data set. Resumes with White

[28] As Appendix Table A1 indicates, the African-American names we use are, however, quite common among African-Americans, making this less of a concern.

[29] In fact, there is some evidence that African-Americans may rely less on social networks for their job search (Harry J. Holzer, 1987).

names have a 9.65 percent chance of receiving a callback. Equivalent resumes with African-American names have a 6.45 percent chance of being called back. This represents a difference in callback rates of 3.20 percentage points, or 50 percent, that can solely be attributed to the name manipulation. Column 4 shows that this difference is statistically significant.[30] Put in other words, these results imply that a White applicant should expect on average one callback for every 10 ads she or he applies to; on the other hand, an African-American applicant would need to apply to about 15 different ads to achieve the same result.[31]

How large are these effects? While the cost of sending additional resumes might not be large per se, this 50-percent gap could be quite substantial when compared to the rate of arrival of new job openings. In our own study, the biggest constraining factor in sending more resumes was the limited number of new job openings each week. Another way to benchmark the measured return to a White name is to compare it to the returns to other resume characteristics. For example, in Table 5, we will show that, at the average number of years of experience in our sample, an extra year of experience increases the likelihood of a callback by a 0.4 percentage point. Based on this point estimate, the return to a White name is equivalent to about eight additional years of experience.

Rows 2 and 3 break down the full sample of sent resumes into the Boston and Chicago markets. About 20 percent more resumes were sent in Chicago than in Boston. The average callback rate (across races) is lower in Chicago than in Boston. This might reflect differences in labor market conditions across the two cities over the experimental period or maybe differences in the ability of the MIT and Chicago teams of research assistants in selecting resumes that were good matches for a given help-wanted ad. The percentage difference in callback rates is, however, strikingly similar across both cities. White applicants are 49 percent more likely than African-American applicants to receive a callback in Chicago and 50 percent more likely in Boston. These racial differences are statistically significant in both cities.

Finally, rows 4 to 7 break down the full sample into female and male applicants. Row 4 displays the average results for all female names while rows 5 and 6 break the female sample into administrative (row 5) and sales jobs (row 6); row 7 displays the average results for all male names. As noted earlier, female names were used in both sales and administrative job openings whereas male names were used close to exclusively for sales openings.[32] Looking across occupations, we find a significant racial gap in callbacks for both males (52 percent) and females (49 percent). Comparing males to females in sales occupations, we find a larger racial gap among males (52 percent versus 22 percent). Interestingly, females in sales jobs appear to receive more callbacks than males; however, this (reverse) gender gap is statistically insignificant and economically much smaller than any of the racial gaps discussed above.

Rather than studying the distribution of callbacks at the applicant level, one can also tabulate the distribution of callbacks at the employment-ad level. In Table 2, we compute the fraction of employers that treat White and African-American applicants equally, the fraction of employers that favor White applicants and the fraction of employers that favor African-American applicants. Because we send up to four resumes in response to each sampled ad, the three categories above can each take three different forms. Equal treatment occurs when either no applicant gets called back, one White and one African-American get called back or two Whites and two African-Americans get called back. Whites are favored when either only one White gets called back, two Whites and no African-American get called back or two Whites and one African-American get called back. African-Americans are favored in all other cases.

As Table 2 indicates, equal treatment occurs for about 88 percent of the help-wanted ads. As expected, the major source of equal treatment comes from the high fraction of ads for which

[30] These statistical tests assume independence of callbacks. We have, however, verified that the results stay significant when we assume that the callbacks are correlated either at the employer or first-name level.

[31] This obviously assumes that African-American applicants cannot assess a priori which firms are more likely to treat them more or less favorably.

[32] Only about 6 percent of all male resumes were sent in response to an administrative job opening.

TABLE 2—DISTRIBUTION OF CALLBACKS BY EMPLOYMENT AD

Equal Treatment:	No Callback	1W + 1B	2W + 2B
88.13 percent	83.37	3.48	1.28
[1,166]	[1,103]	[46]	[17]
Whites Favored (WF):	1W + 0B	2W + 0B	2W + 1B
8.39 percent	5.59	1.44	1.36
[111]	[74]	[19]	[18]
African-Americans Favored (BF):	1B + 0W	2B + 0W	2B + 1W
3.48 percent	2.49	0.45	0.53
[46]	[33]	[6]	[7]
Ho: *WF* = *BF*			
$p = 0.0000$			

Notes: This table documents the distribution of callbacks at the employment-ad level. "No Callback" is the percent of ads for which none of the fictitious applicants received a callback. "1W + 1B" is the percent of ads for which exactly one White and one African-American applicant received a callback. "2W + 2B" is the percent of ads for which exactly two White applicants and two African-American applicants received a callback. "Equal Treatment" is defined as the sum of "No Callback," "1W + 1B," and "2W + 2B." "1W + 0B" is the percent of ads for which exactly one White applicant and no African-American applicant received a call back. "2W + 0B" is the percent of ads for which excatly two White applicants and no African-American applicant received a callback. "2W + 1B" is the percent of ads for which exactly two White applicants and one African-American applicant received a callback. "Whites Favored" is defined as the sum of "1W + 0B," "2W + 0B," and "2W + 1B." "1B + 0W" is the percent of ads for which exactly one African-American applicant and no White applicant received a callback. "2B + 0W" is the percent of ads for which exactly two African-American applicants and no White applicant received a callback. "2B + 1W" is the percent of ads for which exactly two African-American applicants and one White applicant received a callback. "African-Americans Favored" is defined as the sum of "1B + 0W," "2B + 0W," and "2B + 1W." In brackets in each cell is the number of employment ads in that cell. "*Ho*: *WF* = *WB*" reports the *p*-value for a test of symmetry between the proportion of employers that favor White names and the proportion of employers that favor African-American names.

no callbacks are recorded (83 percent of the ads). Whites are favored by nearly 8.4 percent of the employers, with a majority of these employers contacting exactly one White applicant. African-Americans, on the other hand, are favored by only about 3.5 percent of employers. We formally test whether there is symmetry in the favoring of Whites over African-Americans and African-Americans over Whites. We find that the difference between the fraction of employers favoring Whites and the fraction of employers favoring African-Americans is statistically very significant ($p = 0.0000$).

B. *Do African-Americans Receive Different Returns to Resume Quality?*

Our results so far demonstrate a substantial gap in callback based on applicants' names. Next, we would like to learn more about the factors that may influence this gap. More specifically, we ask how employers respond to improvements in African-American applicants' credentials. To answer this question, we examine how the racial gap in callback varies by resume quality.

As we explained in Section II, for most of the employment ads we respond to, we send four different resumes: two higher-quality and two lower-quality ones. Table 3 gives a better sense of which factors enter into this subjective classification. Table 3 displays means and standard deviations of the most relevant resume characteristics for the full sample (column 1), as well as broken down by race (columns 2 and 3) and resume quality (columns 4 and 5). Since applicants' names are randomized, there is no difference in resume characteristics by race. Columns 4 and 5 document the objective differences between resumes subjectively classified as high and low quality. Higher-quality applicants have on average close to an extra year of labor market experience, fewer employment holes (where an employment hole is defined as a period of at least six months without a reported job), are more likely to have worked while at school, and to report some military experience. Also, higher-quality applicants are more likely to have an e-mail address, to have received some honors, and to list some computer skills and other special skills (such as a certification degree or foreign language skills) on their resume. Note that the higher- and lower-quality resumes do not differ on average with regard to

TABLE 3—RESUME CHARACTERISTICS: SUMMARY STATISTICS

Sample:	All resumes	White names	African-American	Higher quality	Lower quality
Characteristic:					
College degree	0.72	0.72	0.72	0.72	0.71
(Y = 1)	(0.45)	(0.45)	(0.45)	(0.45)	(0.45)
Years of experience	7.84	7.86	7.83	8.29	7.39
	(5.04)	(5.07)	(5.01)	(5.29)	(4.75)
Volunteering experience?	0.41	0.41	0.41	0.79	0.03
(Y = 1)	(0.49)	(0.49)	(0.49)	(0.41)	(0.16)
Military experience?	0.10	0.09	0.10	0.19	0.00
(Y = 1)	(0.30)	(0.29)	(0.30)	(0.39)	(0.06)
E-mail address?	0.48	0.48	0.48	0.92	0.03
(Y = 1)	(0.50)	(0.50)	(0.50)	(0.27)	(0.17)
Employment holes?	0.45	0.45	0.45	0.34	0.56
(Y = 1)	(0.50)	(0.50)	(0.50)	(0.47)	(0.50)
Work in school?	0.56	0.56	0.56	0.72	0.40
(Y = 1)	(0.50)	(0.50)	(0.50)	(0.45)	(0.49)
Honors?	0.05	0.05	0.05	0.07	0.03
(Y = 1)	(0.22)	(0.23)	(0.22)	(0.25)	(0.18)
Computer skills?	0.82	0.81	0.83	0.91	0.73
(Y = 1)	(0.38)	(0.39)	(0.37)	(0.29)	(0.44)
Special skills?	0.33	0.33	0.33	0.36	0.30
(Y = 1)	(0.47)	(0.47)	(0.47)	(0.48)	(0.46)
Fraction high school dropouts in applicant's zip code	0.19	0.19	0.19	0.19	0.18
	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)
Fraction college or more in applicant's zip code	0.21	0.21	0.21	0.21	0.22
	(0.17)	(0.17)	(0.17)	(0.17)	(0.17)
Fraction Whites in applicant's zip code	0.54	0.55	0.54	0.53	0.55
	(0.33)	(0.33)	(0.33)	(0.33)	(0.33)
Fraction African-Americans in applicant's zip code	0.31	0.31	0.31	0.32	0.31
	(0.33)	(0.33)	(0.33)	(0.33)	(0.33)
Log(median per capital income) in applicant's zip code	9.55	9.55	9.55	9.54	9.56
	(0.56)	(0.56)	(0.55)	(0.54)	(0.57)
Sample size	4,870	2,435	2,435	2,446	2,424

Notes: The table reports means and standard deviations for the resume characteristics as listed on the left. Column 1 refers to all resumes sent; column 2 refers to resumes with White names; column 3 refers to resumes with African-American names; column 4 refers to higher-quality resumes; column 5 refers to lower-quality resumes. See text for details.

applicants' education level. This reflects the fact that all sent resumes, whether high or low quality, are chosen to be good matches for a given job opening. About 70 percent of the sent resumes report a college degree.[33]

The last five rows of Table 3 show summary characteristics of the applicants' zip code address. Using 1990 Census data, we compute the fraction of high school dropouts, fraction of college educated or more, fraction of Whites, fraction of African-Americans and log(median per capital income) for each zip code used in the experiment. Since addresses are randomized within cities, these neighborhood quality measures are uncorrelated with race or resume quality.

The differences in callback rates between high- and low-quality resumes are presented in Panel A of Table 4. The first thing to note is that the resume quality manipulation works: higher-quality resumes receive more callbacks. As row 1 indicates, we record a callback rate of close to 11 percent for White applicants with a higher-quality resume, compared to 8.5 percent for White applicants with lower-quality resumes. This is a statistically significant difference of 2.29 percentage points, or 27 percent ($p = 0.0557$). Most strikingly, African-Americans experience much less of an increase in callback

[33] This varies from about 50 percent for the clerical and administrative support positions to more than 80 percent for the executive, managerial, and sales representatives positions.

TABLE 4—AVERAGE CALLBACK RATES BY RACIAL SOUNDINGNESS OF NAMES AND RESUME QUALITY

	Panel A: Subjective Measure of Quality (Percent Callback)			
	Low	High	Ratio	Difference (*p*-value)
White names	8.50	10.79	1.27	2.29
	[1,212]	[1,223]		(0.0557)
African-American names	6.19	6.70	1.08	0.51
	[1,212]	[1,223]		(0.6084)
	Panel B: Predicted Measure of Quality (Percent Callback)			
	Low	High	Ratio	Difference (*p*- value)
White names	7.18	13.60	1.89	6.42
	[822]	[816]		(0.0000)
African-American names	5.37	8.60	1.60	3.23
	[819]	[814]		(0.0104)

Notes: Panel A reports the mean callback percents for applicant with a White name (row 1) and African-American name (row 2) depending on whether the resume was subjectively qualified as a lower quality or higher quality. In brackets is the number of resumes sent for each race/quality group. The last column reports the *p*-value of a test of proportion testing the null hypothesis that the callback rates are equal across quality groups within each racial group. For Panel B, we use a third of the sample to estimate a probit regression of the callback dummy on the set of resume characteristics as displayed in Table 3. We further control for a sex dummy, a city dummy, six occupation dummies, and a vector of dummy variables for job requirements as listed in the employment ad (see Section III, subsection D, for details). We then use the estimated coefficients on the set of resume characteristics to estimate a predicted callback for the remaining resumes (two-thirds of the sample). We call "high-quality" resumes the resumes that rank above the median predicted callback and "low-quality" resumes the resumes that rank below the median predicted callback. In brackets is the number of resumes sent for each race/quality group. The last column reports the *p*-value of a test of proportion testing the null hypothesis that the callback percents are equal across quality groups within each racial group.

rate for similar improvements in their credentials. African-Americans with higher-quality resumes receive a callback 6.7 percent of the time, compared to 6.2 percent for African-Americans with lower quality resumes. This is only a 0.51-percentage-point, or 8-percent, difference and this difference is not statistically significant ($p = 0.6084$).

Instead of relying on the subjective quality classification, Panel B directly uses resume characteristics to classify the resumes. More specifically, we use a random subsample of one-third of the resumes to estimate a probit regression of the callback dummy on the resume characteristics listed in Table 3. We further control for a sex dummy, a city dummy, six occupation dummies, and a vector of job requirements as listed in the employment ads.[34] We then use the estimated coefficients on the resume characteristics to rank the remaining two-thirds of the resumes by predicted callback. In Panel B, we classify as "high" those resumes that have above-median-predicted callback; similarly, we classify as "low" those resumes that have below-median-predicted callback. As one can see from Panel B, qualitatively similar results emerge from this analysis. While African-Americans do appear to significantly benefit from higher-quality resumes under this alternative classification, they benefit less than Whites. The ratio of callback rates for high- versus low-quality resumes is 1.60 for African Americans, compared to 1.89 for Whites.

In Table 5, we directly report the results of race-specific probit regressions of the callback dummy on resume characteristics. We, however, start in column 1 with results for the full sample of sent resumes. As one can see, many of the resume characteristics have the expected effect on the likelihood of a callback. The addition of an e-mail address, honors, and special skills all have a positive and significant effect on the likelihood of a callback.[35] Also, more experienced applicants are more likely to get called back: at the average number of years of experience in our sample (eight years), each

[34] See Section III, subsection D, for more details on these occupation categories and job requirements.

[35] Note that the e-mail address dummy, because it is close to perfectly correlated with the subjective resume-quality variable, may in part capture some other unmeasured resume characteristics that may have led us to categorize a given resume as higher quality.

TABLE 5—EFFECT OF RESUME CHARACTERISTICS ON LIKELIHOOD OF CALLBACK

Dependent Variable: Callback Dummy Sample:	All resumes	White names	African-American names
Years of experience (*10)	0.07	0.13	0.02
	(0.03)	(0.04)	(0.03)
Years of experience^2 (*100)	−0.02	−0.04	−0.00
	(0.01)	(0.01)	(0.01)
Volunteering? (Y = 1)	−0.01	−0.01	0.01
	(0.01)	(0.01)	(0.01)
Military experience? (Y = 1)	−0.00	0.02	−0.01
	(0.01)	(0.03)	(0.02)
E-mail? (Y = 1)	0.02	0.03	−0.00
	(0.01)	(0.01)	(0.01)
Employment holes? (Y = 1)	0.02	0.03	0.01
	(0.01)	(0.02)	(0.01)
Work in school? (Y = 1)	0.01	0.02	−0.00
	(0.01)	(0.01)	(0.01)
Honors? (Y = 1)	0.05	0.06	0.03
	(0.02)	(0.03)	(0.02)
Computer skills? (Y = 1)	−0.02	−0.04	−0.00
	(0.01)	(0.02)	(0.01)
Special skills? (Y = 1)	0.05	0.06	0.04
	(0.01)	(0.02)	(0.01)
Ho: Resume characteristics effects are all zero (*p*-value)	54.50 (0.0000)	57.59 (0.0000)	23.85 (0.0080)
Standard deviation of predicted callback	0.047	0.062	0.037
Sample size	4,870	2,435	2,435

Notes: Each column gives the results of a probit regression where the dependent variable is the callback dummy. Reported in the table are estimated marginal changes in probability for the continuous variables and estimated discrete changes for the dummy variables. Also included in each regression are a city dummy, a sex dummy, six occupation dummies, and a vector of dummy variables for job requirements as listed in the employment ad (see Section III, subsection D, for details). Sample in column 1 is the entire set of sent resumes; sample in column 2 is the set of resumes with White names; sample in column 3 is the set of resumes with African-American names. Standard errors are corrected for clustering of the observations at the employment-ad level. Reported in the second to last row are the *p*-values for a χ^2 testing that the effects on the resume characteristics are all zero. Reported in the second to last row is the standard deviation of the predicted callback rate.

extra year of experience increases the likelihood of a callback by about a 0.4 percentage point. The most counterintuitive effects come from computer skills, which appear to negatively predict callback, and employment holes, which appear to positively predict callback.

The same qualitative patterns hold in column 2 where we focus on White applicants. More importantly, the estimated returns to an e-mail address, additional work experience, honors, and special skills appear economically stronger for that racial group. For example, at the average number of years of experience in our sample, each extra year of experience increases the likelihood of a callback by about a 0.7 percentage point.

As might have been expected from the two previous columns, we find that the estimated returns on these resume characteristics are all economically and statistically weaker for African-American applicants (column 3). In fact, all the estimated effects for African-Americans are statistically insignificant, except for the return to special skills. Resume characteristics thus appear less predictive of callback rates for African-Americans than they are for Whites. To illustrate this more saliently, we predict callback rates using either regression estimates in column 2 or regression estimates in column 3. The standard deviation of the predicted callback from column 2 is 0.062, whereas it is only 0.037 from column 3. In summary, employers simply seem to pay less attention or discount more the characteristics listed on the

TABLE 6—EFFECT OF APPLICANT'S ADDRESS ON LIKELIHOOD OF CALLBACK

Dependent Variable: Callback Dummy						
Zip code characteristic:	Fraction Whites		Fraction college or more		Log(per capital income)	
Zip code characteristic	0.020	0.020	0.054	0.053	0.018	0.014
	(0.012)	(0.016)	(0.022)	(0.031)	(0.007)	(0.010)
Zip code characteristic* African-American name	—	−0.000	—	−0.002	—	0.008
		(0.024)		(0.048)		(0.015)
African-American name	—	−0.031	—	−0.031	—	−0.112
		(0.015)		(0.013)		(0.152)

Notes: Each column gives the results of a probit regression where the dependent variable is the callback dummy. Reported in the table is the estimated marginal change in probability. Also included in columns 1, 3, and 5 is a city dummy; also included in columns 2, 4, and 6 is a city dummy and a city dummy interacted with a race dummy. Standard errors are corrected for clustering of the observations at the employment-ad level.

resumes with African-American-sounding names. Taken at face value, these results suggest that African-Americans may face relatively lower individual incentives to invest in higher skills.[36]

C. *Applicants' Address*

An incidental feature of our experimental design is the random assignment of addresses to the resumes. This allows us to examine whether and how an applicant's residential address, all else equal, affects the likelihood of a callback. In addition, and most importantly for our purpose, we can also ask whether African-American applicants are helped relatively more by residing in more affluent neighborhoods.

We perform this analysis in Table 6. We start (columns 1, 3, and 5) by discussing the effect of neighborhood of residence across all applicants. Each of these columns reports the results of a probit regression of the callback dummy on a specific zip code characteristic and a city dummy. Standard errors are corrected for clustering of the observations at the employment-ad level. We find a positive and significant effect of neighborhood quality on the likelihood of a callback. Applicants living in Whiter (column 1), more educated (column 3), or higher-income (column 5) neighborhoods have a higher probability of receiving a callback. For example, a 10-percentage-point increase in the fraction of college-educated in zip code of residence increases the likelihood of a callback by a 0.54 percentage point (column 3).

In columns 2, 4, and 6, we further interact the zip code characteristic with a dummy variable for whether the applicant is African-American or not. Each of the probit regressions in these columns also includes an African-American dummy, a city dummy, and an interaction of the city dummy with the African-American dummy. There is no evidence that African-Americans benefit any more than Whites from living in a Whiter, more educated zip code. The estimated interactions between fraction White and fraction college educated with the African-American dummy are economically very small and statistically insignificant. We do find an economically more meaningful effect of zip code median income level on the racial gap in callback; this effect, however, is statistically insignificant.

In summary, while neighborhood quality affects callbacks, African-Americans do not benefit more than Whites from living in better neighborhoods. If ghettos and bad neighborhoods are particularly stigmatizing for African-Americans, one might have expected African-Americans to be helped more by having a "better" address. Our results do not support this hypothesis.

D. *Job and Employer Characteristics*

Table 7 studies how various job requirements (as listed in the employment ads) and employer characteristics correlate with the racial gap in callback. Each row of Table 7 focuses on a specific job or employer characteristic, with

[36] This of course assumes that the changes in job and wage offers associated with higher skills are the same across races, or at least not systematically larger for African-Americans.

TABLE 7—EFFECT OF JOB REQUIREMENT AND EMPLOYER CHARACTERISTICS ON RACIAL DIFFERENCES IN CALLBACKS

Job requirement:	Sample mean (standard deviation)	Marginal effect on callbacks for African-American names
Any requirement? (Y = 1)	0.79	0.023
	(0.41)	(0.015)
Experience? (Y = 1)	0.44	0.011
	(0.49)	(0.013)
Computer skills? (Y = 1)	0.44	0.000
	(0.50)	(0.013)
Communication skills? (Y = 1)	0.12	−0.000
	(0.33)	(0.015)
Organization skills? (Y = 1)	0.07	0.028
	(0.26)	(0.029)
Education? (Y = 1)	0.11	−0.031
	(0.31)	(0.017)
Total number of requirements	1.18	0.002
	(0.93)	(0.006)
Employer characteristic:	**Sample mean (standard deviation)**	**Marginal effect on callbacks for African-American names**
Equal opportunity employer? (Y = 1)	0.29	−0.013
	(0.45)	(0.012)
Federal contractor? (Y = 1)	0.11	−0.035
(N = 3,102)	(0.32)	(0.016)
Log(employment)	5.74	−0.001
(N = 1,690)	(1.74)	(0.005)
Ownership status:		
(N = 2,878)		
Privately held	0.74	0.011
		(0.019)
Publicly traded	0.15	−0.025
		(0.015)
Not-for-profit	0.11	0.025
		(0.042)
Fraction African-Americans in employer's zip code	0.08	0.117
(N = 1,918)	(0.15)	(0.062)

Notes: Sample is all sent resumes (N = 4,870) unless otherwise specified in column 1. Column 2 reports means and standard deviations (in parentheses) for the job requirement or employer characteristic. For ads listing an experience requirement, 50.1 percent listed "some," 24.0 percent listed "two years or less," and 25.9 percent listed "three years or more." For ads listing an education requirement, 8.8 percent listed a high school degree, 48.5 percent listed some college, and 42.7 percent listed at least a four-year college degree. Column 3 reports the marginal effect of the job requirement or employer characteristic listed in that row on differential treatment. Specifically, each cell in column 3 corresponds to a different probit regression of the callback dummy on an African-American name dummy, a dummy for the requirement or characteristic listed in that row and the interaction of the requirement or characteristic dummy with the African-American name dummy. Reported in each cell is the estimated change in probability for the interaction term. Standard errors are corrected for clustering of the observations at the employment-ad level.

summary statistics in column 2. Column 3 shows the results of various probit regressions. Each entry in this column is the marginal effect of the specific characteristic listed in that row on the racial gap in callback. More specifically, each entry is from a separate probit regression of a callback dummy on an African-American dummy, the characteristic listed in that row and the interaction of that characteristic with the African-American dummy. The reported coefficient is that on the interaction term.

We start with job requirements. About 80 percent of the ads state some form of requirement. About 44 percent of the ads require some minimum experience, of which roughly 50 percent simply ask for "some experience," 24 percent less than two years, and 26 percent at least three years of experience. About 44 percent of

ads mention some computer knowledge requirement, which can range from Excel or Word to more esoteric software programs. Good communication skills are explicitly required in about 12 percent of the ads. Organization skills are mentioned 7 percent of the time. Finally, only about 11 percent of the ads list an explicit education requirement. Of these, 8.8 percent require a high school degree, 48.5 percent some college (such as an associate degree), and the rest at least a four-year college degree.[37]

Despite this variability, we find little systematic relationship between any of the requirements and the racial gap in callback. The point estimates in column 3 show no consistent economic pattern and are all statistically weak. Measures of job quality, such as experience or computer skills requirements, do not predict the extent of the racial gap. Communication or other interpersonal skill requirements have no effect on the racial gap either.[38]

We also study employer characteristics. Collecting such information is a more difficult task since it is not readily available from the employment ads we respond to. The only piece of employer information we can directly collect from the employment ad is whether or not the employer explicitly states being an "Equal Opportunity Employer." In several cases, the name of the employer is not even mentioned in the ad and the only piece of information we can rely on is the fax number which applications must be submitted to. We therefore have to turn to supplemental data sources. For employment ads that do not list a specific employer, we first use the fax number to try to identify the company name via Web reverse-lookup services. Based on company names, we use three different data sources (*Onesource Business Browser, Thomas Register,* and *Dun and Bradstreet Million Dollar Directory, 2001*) to track company information such as total employment, industry, and ownership status. Using this same set of data sources, we also try to identify the specific zip code of the company (or company branch) that resumes are to be sent to. Finally, we use the Federal Procurement and Data Center Web site to find a list of companies that have federal contracts.[39] The racial difference in callback rates for the subsamples where employer characteristics could be determined is very similar in magnitude to that in the full sample.

Employer characteristics differ significantly across ads. Twenty-nine percent of all employers explicitly state that they are "Equal Opportunity Employers." Eleven percent are federal contractors and, therefore, might face greater scrutiny under affirmative action laws. The average company size is around 2,000 employees but there is a lot of variation across firms. Finally, 74 percent of the firms are privately held, 15 percent are publicly traded, and 11 percent are not-for-profit organizations.

Neither "Equal Opportunity Employers" nor federal contractors appear to treat African-Americans more favorably. In fact, each of these employer characteristics is associated with a larger racial gap in callback (and this effect is marginally significant for federal contractors). Differential treatment does not vary with employer size.[40] Point estimates indicate less differential treatment in the not-for-profit sector; however, this effect is very noisily estimated.[41]

In an unpublished Appendix (available from the authors upon request), we also study how the racial gap in callback varies by occupation and industry. Based on the employment ad listings, we classify the job openings into six occupation categories: executives and managers; administrative supervisors; sales representatives; sales workers; secretaries and legal assistants; clerical workers. We also, when possible,

[37] Other requirements sometimes mentioned include typing skills for secretaries (with specific words-per-minute minimum thresholds), and, more rarely, foreign language skills.

[38] Other ways of estimating these effects produce a similar nonresult. Among other things, we considered including a city dummy or estimating the effects separately by city; we also estimated one single probit regression including all requirements at once.

[39] This Web site (www.fpdc.gov) is accurate up to and including March 21, 2000.

[40] Similar results hold when we measure employer size using a total sales measure rather than an employment measure.

[41] Our measurement of the racial gap by firm or employer type may not be a good indicator of the fraction of African-Americans actually employed in these firms. For example, "Equal Opportunity Employers" may receive a higher fraction of African-American resumes. Their actual hiring may therefore look different from that of non "Equal Opportunity Employers" when one considers the full set of resumes they receive.

classify employers into six industry categories: manufacturing; transportation and communication; wholesale and retail trade; finance, insurance, and real estate; business and personal services; health, educational, and social services. We then compute occupation and industry-specific racial gaps in callback and relate these gaps to 1990 Census-based measures of occupation and industry earnings, as well as Census-based measures of the White/African-American wage gap in these occupations and industries.

We find a positive White/African-American gap in callbacks in all occupation and industry categories (except for transportation and communication). While average earnings vary a lot across the occupations covered in the experiment, we find no systematic relationship between occupation earnings and the racial gap in callback. Similarly, the industry-specific gaps in callback do not relate well to a measure of inter-industry wage differentials. In fact, while the racial gap in callback rates varies somewhat across occupations and industries, we cannot reject the null hypothesis that the gap is the same across all these categories.

The last row of Table 7 focuses on the marginal effect of employer location on the racial gap in callback.[42] We use as a measure of employer location the zip code of the company (or company branch) resumes were to be sent to. More specifically, we ask whether differential treatment by race varies with the fraction of African-Americans in the employer's zip code. We find a marginally significant positive effect of employer location on African-American callbacks but this effect is extremely small. In regressions not reported here (but available from the authors upon request), we reestimate this effect separately by city. While the point estimates are positive for both cities, the effect is only statistically significant for Chicago.

IV. Interpretation

Three main sets of questions arise when interpreting the results above. First, does a higher callback rate for White applicants imply that employers are discriminating against African-Americans? Second, does our design only isolate the effect of race or is the name manipulation conveying some other factors than race? Third, how do our results relate to different models of racial discrimination?

A. *Interpreting Callback Rates*

Our results indicate that for two identical individuals engaging in an identical job search, the one with an African-American name would receive fewer interviews. Does differential treatment within our experiment imply that employers are discriminating against African-Americans (whether it is rational, prejudice-based, or other form of discrimination)? In other words, could the lower callback rate we record for African-American resumes *within our experiment* be consistent with a racially neutral review of the *entire pool* of resumes the surveyed employers receive?

In a racially neutral review process, employers would rank order resumes based on their quality and call back all applicants that are above a certain threshold. Because names are randomized, the White and African-American resumes we send should rank similarly on average. So, irrespective of the skill and racial composition of the applicant pool, a race-blind selection rule would generate equal treatment of Whites and African-Americans. So our results must imply that employers use race as a factor when reviewing resumes, which matches the legal definition of discrimination.

But even rules where employers are not trying to interview as few African-American applicants as possible may generate observed differential treatment in our experiment. One such hiring rule would be employers trying to interview a target level of African-American candidates. For example, perhaps the average firm in our experiment aims to produce an interview pool that matches the population base rate. This rule could produce the observed differential treatment if the average firm receives a higher proportion of African-American resumes than the population base rate because African-Americans disproportionately apply to the jobs and industries in our sample.[43]

[42] For previous work on the effect of employer location on labor market discrimination, see, for example, Steven Raphael et al. (2000).

[43] Another variant of this argument is that the (up to) two African-American resumes we sent are enough to signifi-

Some of our other findings may be consistent with such a rule. For example, the fact that "Equal Opportunity Employers" or federal contractors do not appear to discriminate any less may reflect the fact that such employers receive more applications from African-Americans. On the other hand, other key findings run counter to this rule. As we discuss above, we find no systematic difference in the racial gap in callback across occupational or industry categories, despite the large variation in the fraction of African-Americans looking for work in those categories. African-Americans are underrepresented in managerial occupations, for example. If employers matched base rates in the population, the few African-Americans who apply to these jobs should receive a higher callback rate than Whites. Yet, we find that the racial gap in managerial occupations is the same as in all the other job categories. This rule also runs counter to our findings on returns to skill. Suppose firms are struggling to find White applicants but overwhelmed with African-American ones. Then they should be less sensitive to the quality of White applicants (as they are trying to fill in their hiring quota for Whites) and much more sensitive to the quality of Black applicants (when they have so many to pick from). Thus, it is unlikely that the differential treatment we observe is generated by hiring rules such as these.

cantly distort the racial composition of the entire applicant pool. This is unlikely for two reasons. First, anecdotal evidence and the empirically low callback rates we record suggest that firms typically receive many hundreds of resumes in response to each ad they post. Hence, the (up to) four resumes we send out are unlikely to influence the racial composition of the pool. Second, the similar racial gap in callback we observe across the two cities goes counter to this interpretation since the racial composition base rates differ quite a lot across these two cities. Another variant of this argument is that, for some reason, the average firm in our sample receives a lot of high-quality resumes from African-American applicants and much fewer high-quality resumes from White applicants. Hypothetically, this might occur if high-quality African-Americans are much more likely to use help-wanted ads rather than other job search channels. If employers perform within-race comparisons and again want to target a certain racial mix in their interviewing and hiring, our African-American resumes may naturally receive lower callbacks as they are competing with many more high-quality applicants. This specific argument would be especially relevant in a case where the average sampled employer is "known" to be good to African-Americans. But our selection procedure for the employment ads did not allow for such screening: we simply responded to as many ads as possible in the targeted occupational categories.

B. *Potential Confounds*

While the names we have used in this experiment strongly signal racial origin, they may also signal some other personal trait. More specifically, one might be concerned that employers are inferring social background from the personal name. When employers read a name like "Tyrone" or "Latoya," they may assume that the person comes from a disadvantaged background.[44] In the extreme form of this social background interpretation, employers do not care at all about race but are discriminating only against the social background conveyed by the names we have chosen.[45]

While plausible, we feel that some of our earlier results are hard to reconcile with this interpretation. For example, in Table 6, we found that while employers value "better" addresses, African-Americans are not helped more than Whites by living in Whiter or more educated neighborhoods. If the African-American names we have chosen mainly signal negative social background, one might have expected the estimated name gap to be lower for better addresses. Also, if the names mainly signal social background, one might have expected the name gap to be higher for jobs that rely more on soft skills or require more interpersonal interactions. We found no such evidence in Table 7.

We, however, directly address this alternative interpretation by examining the average social background of babies born with the names used in the experiment. We were able to obtain birth certificate data on mother's education (less than high school, high school or more) for babies born in Massachusetts between 1970 and

[44] Roland Fryer and Steven Levitt (2003) provide a recent analysis of social background and naming conventions amongst African-Americans.

[45] African-Americans as a whole come from more disadvantaged backgrounds than Whites. For this social class effect to be something of independent interest, one must assert that African-Americans with the African-American names we have selected are from a lower social background than the average African-American and/or that Whites with the White names we have selected are from a higher social background than the average White. We come back to this point below.

TABLE 8—CALLBACK RATE AND MOTHER'S EDUCATION BY FIRST NAME

White female			African-American female		
Name	Percent callback	Mother education	Name	Percent callback	Mother education
Emily	7.9	96.6	Aisha	2.2	77.2
Anne	8.3	93.1	Keisha	3.8	68.8
Jill	8.4	92.3	Tamika	5.5	61.5
Allison	9.5	95.7	Lakisha	5.5	55.6
Laurie	9.7	93.4	Tanisha	5.8	64.0
Sarah	9.8	97.9	Latoya	8.4	55.5
Meredith	10.2	81.8	Kenya	8.7	70.2
Carrie	13.1	80.7	Latonya	9.1	31.3
Kristen	13.1	93.4	Ebony	9.6	65.6
Average		91.7	Average		61.0
Overall		83.9	Overall		70.2
Correlation	−0.318	($p = 0.404$)	Correlation	−0.383	($p = 0.309$)
White male			African-American male		
Name	Percent callback	Mother education	Name	Percent callback	Mother education
Todd	5.9	87.7	Rasheed	3.0	77.3
Neil	6.6	85.7	Tremayne	4.3	—
Geoffrey	6.8	96.0	Kareem	4.7	67.4
Brett	6.8	93.9	Darnell	4.8	66.1
Brendan	7.7	96.7	Tyrone	5.3	64.0
Greg	7.8	88.3	Hakim	5.5	73.7
Matthew	9.0	93.1	Jamal	6.6	73.9
Jay	13.4	85.4	Leroy	9.4	53.3
Brad	15.9	90.5	Jermaine	9.6	57.5
Average		91.7	Average		66.7
Overall		83.5	Overall		68.9
Correlation	−0.0251	($p = 0.949$)	Correlation	−0.595	($p = 0.120$)

Notes: This table reports, for each first name used in the experiment, callback rate and average mother education. Mother education for a given first name is defined as the percent of babies born with that name in Massachusetts between 1970 and 1986 whose mother had at least completed a high school degree (see text for details). Within each sex/race group, first names are ranked by increasing callback rate. "Average" reports, within each race-gender group, the average mother education for all the babies born with one of the names used in the experiment. "Overall" reports, within each race-gender group, average mother education for all babies born in Massachusetts between 1970 and 1986 in that race-gender group. "Correlation" reports the Spearman rank order correlation between callback rate and mother education *within* each race-gender group as well as the *p*-value for the test of independence.

1986.[46] For each first name in our experiment, we compute the fraction of babies with that name and, in that gender-race cell, whose mothers have at least completed a high school degree.

In Table 8, we display the average callback rate for each first name along with this proxy for social background. Within each race-gender group, the names are ranked by increasing callback rate. Interestingly, there is significant

[46] This longer time span (compared to that used to assess name frequencies) was imposed on us for confidentiality reasons. When fewer than 10 births with education data available are recorded in a particular education-name cell, the exact number of births in that cell is not reported and we impute five births. Our results are not sensitive to this imputation. One African-American female name (Latonya) and two male names (Rasheed and Hakim) were imputed in this way. One African-American male name (Tremayne) had too few births with available education data and was therefore dropped from this analysis. Our results are qualitatively similar when we use a larger data set of California births for the years 1989 to 2000 (kindly provided to us by Steven Levitt).

variation in callback rates by name. Of course, chance alone could produce such variation because of the rather small number of observations in each cell (about 200 for the female names and 70 for the male names).[47]

The row labeled "Average" reports the average fraction of mothers that have at least completed high school for the set of names listed in that gender-race group. The row labeled "Overall" reports the average fraction of mothers that have at least completed high school for the full sample of births in that gender-race group. For example, 83.9 percent of White female babies born between 1970 and 1986 have mothers with at least a high school degree; 91.7 percent of the White female babies with one of the names used in the experiment have mothers with at least a high school degree.

Consistent with a social background interpretation, the African-American names we have chosen fall below the African-American average. For African-American male names, however, the gap between the experimental names and the population average is negligible. For White names, both the male and female names are above the population average.

But, more interestingly to us, there is substantial between-name heterogeneity in social background. African-American babies named Kenya or Jamal are affiliated with much higher mothers' education than African-American babies named Latonya or Leroy. Conversely, White babies named Carrie or Neil have lower social background than those named Emily or Geoffrey. This allows for a direct test of the social background hypothesis within our sample: are names associated with a worse social background discriminated against more? In the last row in each gender-race group, we report the rank-order correlation between callback rates and mother's education. The social background hypothesis predicts a positive correlation. Yet, for all four categories, we find the exact opposite. The *p*-values indicate that we cannot reject independence at standard significance levels except in the case of African-American males where we can almost reject it at the 10-percent level ($p = 0.120$). In summary, this test suggests little evidence that social background drives the measured race gap.

Names might also influence our results through familiarity. One could argue that the African-American names used in the experiment simply appear odd to human resource managers and that any odd name is discriminated against. But as noted earlier, the names we have selected are not particularly uncommon among African-Americans (see Appendix Table A1). We have also performed a similar exercise to that of Table 8 and measured the rank-order correlation between name-specific callback rates and name frequency within each gender-race group. We found no systematic positive correlation.

There is one final potential confound to our results. Perhaps what appears as a bias against African-Americans is actually the result of *reverse discrimination.* If qualified African-Americans are thought to be in high demand, then employers with average quality jobs might feel that an equally talented African-American would never accept an offer from them and thereby never call her or him in for an interview. Such an argument might also explain why African-Americans do not receive as strong a return as Whites to better resumes, since higher qualification only strengthens this argument. But this interpretation would suggest that among the better jobs, we ought to see evidence of reverse discrimination, or at least a smaller racial gap. However, as we discussed in Section III, subsection D, we do not find any such evidence. The racial gap does not vary across jobs with different skill requirements, nor does it vary across occupation categories. Even among the better jobs in our sample, we find that employers significantly favor applicants with White names.[48]

[47] We formally tested whether this variation was significant by estimating a probit regression of the callback dummy on all the personal first names, allowing for clustering of the observations at the employment-ad level. For all but African-American females, we cannot reject the null hypothesis that all the first name effects in the same race-gender group are the same. Of course, a lack of a rejection does not mean there is no underlying pattern in the between-name variation in callbacks that might have been detectable with larger sample sizes.

[48] One might argue that employers who reverse-discriminate hire through less formal channels than help-wanted ads. But this would imply that African-Americans are less likely to find jobs through formal channels. The evidence on exit out of unemployment does not paint a clear picture in this direction (Holzer, 1987).

C. *Relation to Existing Theories*

What do these results imply for existing models of discrimination? Economic theories of discrimination can be classified into two main categories: taste-based and statistical discrimination models.[49] Both sets of models can obviously "explain" our average racial gap in callbacks. But can these models explain our other findings? More specifically, we discuss the relevance of these models with a focus on two of the facts that have been uncovered in this paper: (i) the lower returns to credentials for African-Americans; (ii) the relative uniformity of the race gap across occupations, job requirements and, to a lesser extent, employer characteristics and industries.

Taste-based models (Gary S. Becker, 1961) differ in whose prejudiced "tastes" they emphasize: customers, coworkers, or employers. Customer and co-worker discrimination models seem at odds with the lack of significant variation of the racial gap by occupation and industry categories, as the amount of customer contact and the fraction of White employees vary quite a lot across these categories. We do not find a larger racial gap among jobs that explicitly require "communication skills" and jobs for which we expect either customer or coworker contacts to be higher (retail sales for example).

Because we do not know what drives employer tastes, employer discrimination models could be consistent with the lack of occupation and industry variation. Employer discrimination also matches the finding that employers located in more African-American neighborhoods appear to discriminate somewhat less. However, employer discrimination models would struggle to explain why African-Americans get relatively lower returns to their credentials. Indeed, the cost of indulging the discrimination taste should increase as the minority applicants' credentials increase.[50]

Statistical discrimination models are the prominent alternative to the taste-based models in the economics literature. In one class of statistical discrimination models, employers use (observable) race to proxy for *unobservable* skills (e.g., Edmund S. Phelps, 1972; Kenneth J. Arrow, 1973). This class of models struggle to explain the credentials effect as well. Indeed, the added credentials should lead to a larger update for African-Americans and hence greater returns to skills for that group.

A second class of statistical discrimination models "emphasize the precision of the information that employers have about individual productivity" (Altonji and Blank, 1999). Specifically, in these models, employers believe that the same observable signal is more precise for Whites than for African-Americans (Dennis J. Aigner and Glenn G. Cain, 1977; Shelly J. Lundberg and Richard Startz, 1983; Bradford Cornell and Ivo Welch, 1996). Under such models, African-Americans receive lower returns to observable skills because employers place less weight on these skills. However, how reasonable is this interpretation for our experiment? First, it is important to note that we are using the same set of resume characteristics for both racial groups. So the lower precision of information for African-Americans cannot be that, for example, an employer does not know what a high school degree from a very African-American neighborhood means (as in Aigner and Cain, 1977). Second, many of the credentials on the resumes are in fact externally and easily verifiable, such as a certification for a specific software.

An alternative version of these models would rely on bias in the observable signal rather than differential variance or noise of these signals by race. Perhaps the skills of African-Americans are discounted because affirmative action makes it easier for African-Americans to get these skills. While this is plausible for credentials such as an employee-of-the-month honor, it is unclear why this would apply to more verifiable and harder skills. It is equally unclear why work experience would be less rewarded since our study suggests that getting a job is more, not less, difficult for African-Americans.

The uniformity of the racial gap across occupations is also troubling for a statistical discrimination interpretation. Numerous factors that should affect the level of statistical discrimination, such as the importance of unobservable skills, the observability of qualifications, the precision of observable skills and the ease of

[49] Darity and Mason (1998) provide a more thorough review of a variety of economic theories of discrimination.

[50] One could, however, assume that employer tastes differ not just by race but also by race and skill, so that employers have greater prejudice against minority workers with better credentials. But the opposite preferences, employers having a particular distaste for low-skilled African-Americans, also seem reasonable.

performance measurement, may vary quite a lot across occupations.

This discussion suggests that perhaps other models may do a better job at explaining our findings. One simple alternative model is lexicographic search by employers. Employers receive so many resumes that they may use quick heuristics in reading these resumes. One such heuristic could be to simply read no further when they see an African-American name. Thus they may never see the skills of African-American candidates and this could explain why these skills are not rewarded. This might also to some extent explain the uniformity of the race gap since the screening process (i.e., looking through a large set of resumes) may be quite similar across the variety of jobs we study.[51]

[51] Another explanation could be based on employer stereotyping or categorizing. If employers have coarser stereotypes for African-Americans, many of our results would follow. See Melinda Jones (2002) for the relevant psychology and Mullainathan (2003) for a formalization of the categorization concept.

V. Conclusion

This paper suggests that African-Americans face differential treatment when searching for jobs and this may still be a factor in why they do poorly in the labor market. Job applicants with African-American names get far fewer callbacks for each resume they send out. Equally importantly, applicants with African-American names find it hard to overcome this hurdle in callbacks by improving their observable skills or credentials.

Taken at face value, our results on differential returns to skill have possibly important policy implications. They suggest that training programs alone may not be enough to alleviate the racial gap in labor market outcomes. For training to work, some general-equilibrium force outside the context of our experiment would have to be at play. In fact, if African-Americans recognize how employers reward their skills, they may rationally be less willing than Whites to even participate in these programs.

TABLE A1—FIRST NAMES USED IN EXPERIMENT

White female			African-American female		
Name	L(W)/L(B)	Perception White	Name	L(B)/L(W)	Perception Black
Allison	∞	0.926	Aisha	209	0.97
Anne	∞	0.962	Ebony	∞	0.9
Carrie	∞	0.923	Keisha	116	0.93
Emily	∞	0.925	Kenya	∞	0.967
Jill	∞	0.889	Lakisha	∞	0.967
Laurie	∞	0.963	Latonya	∞	1
Kristen	∞	0.963	Latoya	∞	1
Meredith	∞	0.926	Tamika	284	1
Sarah	∞	0.852	Tanisha	∞	1
	Fraction of all births:			Fraction of all births:	
	3.8 percent			7.1 percent	
White male			African-American male		
Name	L(W)/L(B)	Perception White	Name	L(B)/L(W)	Perception Black
Brad	∞	1	Darnell	∞	0.967
Brendan	∞	0.667	Hakim		0.933
Geoffrey	∞	0.731	Jamal	257	0.967
Greg	∞	1	Jermaine	90.5	1
Brett	∞	0.923	Kareem	∞	0.967
Jay	∞	0.926	Leroy	44.5	0.933
Matthew	∞	0.888	Rasheed	∞	0.931
Neil	∞	0.654	Tremayne	∞	0.897
Todd	∞	0.926	Tyrone	62.5	0.900
	Fraction of all births:			Fraction of all births:	
	1.7 percent			3.1 percent	

Notes: This table tabulates the different first names used in the experiment and their identifiability. The first column reports the likelihood that a baby born with that name (in Massachusetts between 1974 and 1979) is White (or African-American) relative to the likelihood that it is African-American (White). The second column reports the probability that the name was picked as White (or African-American) in an independent field survey of people. The last row for each group of names shows the proportion of all births in that race group that these names account for.

REFERENCES

Aigner, Dennis J. and Cain. Glenn G. "Statistical Theories of Discrimination in Labor Markets." *Industrial and Labor Relations Review*, January 1977, *30*(1), pp. 175–87.

Altonji, Joseph G. and Blank, Rebecca M. "Race and Gender in the Labor Markey," in Orley Ashenfelter and David Card, eds., *Handbook of labor economics*, Vol. 30. Amsterdam: North-Holland, 1999, pp. 3143–259.

Arrow, Kenneth, J. "The Theory of Discrimination," in Orley Ashenfelter and Albert Rees, eds., *Discrimination in labor markets*. Princeton, NJ: Princeton University Press, 1973, pp. 3–33.

_____. "What Has Economics to Say about Racial Discrimination?" *Journal of Economic Perspectives*, Spring 1998, *12*(2), pp. 91–100.

Becker, Gary S. *The economics of discrimination*, 2nd Ed. Chicago: University of Chicago Press, 1961.

Brown, Colin and Gay, Pat. *Racial discrimination 17 years after the act.* London: Policy Studies Institute, 1985.

Cornell, Bradford and Welch, Ivo. "Culture, Information, and Screening Discrimination." *Journal of Political Economy*, June 1996, *104*(3), pp. 542–71.

Council of Economic Advisers. *Changing America: Indicators of social and economic well-being by race and Hispanic origin.* September 1998, http://w3.access.gpo.gov/eop/ca/pdfs/ca.pdf.

Cross, Harry; Kenney, Genevieve; Mell, Jane and Zimmerman, Wendy. *Employer hiring practices: Differential treatment of Hispanic and Anglo job applicants.* Washington, DC: Urban Institute Press, 1990.

Darity, William A., Jr. and Mason, Patrick L. "Evidence on Discrimination in Employment: Codes of Color, Codes of Gender." *Journal of Economic Perspectives,* Spring 1998, *12*(2), pp. 63–90.

Fix, Michael and Turner, Margery A., eds. *A national report card on discrimination in America: The role of testing.* Washington, DC: Urban Institute Press, 1998.

Fryer, Roland and Levitt, Steven. "The Causes and Consequences of Distinctively Black Names." Mimeo, University of Chicago, 2003.

Goldin, Claudia and Rouse, Cecilia. "Orchestrating Impartiality: The Impact of Blind Auditions on Female Musicians." *American Economic Review,* September 2000, *90*(4), pp. 715–41.

Heckman, James J. "Detecting Discrimination." *Journal of Economic Perspectives,* Spring 1998, *12*(2), pp. 101–16.

Heckman, James J.; Lochner, Lance J., and Todd, Petra E. "Fifty Years of Mincer Earnings Regressions." Mimeo, University of Chicago, 2001.

Heckman, James J. and Siegelman, Peter. "The Urban Institute Audit Studies: Their Methods and Findings," in Michael Fix and Raymond J. Struyk, eds., *Clear and convincing evidence: Measurement of discrimination in America.* Lanham, MD: Urban Institute Press, 1992, pp. 187–258.

Holzer, Harry J. "Informal Job Search and Black Youth Unemployment." *American Economic Review,* June 1987, *77*(3), pp. 446–52.

Hubbuck, Jim and Carter, Simon. *Half a chance? A report on job discrimination against young blacks in Nottingham.* London: Commission for Racial Equality, 1980.

James, Franklin and DelCastillo, Steve W. "Measuring Job Discrimination by Private Employers Against Young Black and Hispanic Seeking Entry Level Work in the Denver Metropolitan Area." Mimeo, University of Colorado-Denver, 1991.

Jones, Melinda. *Social psychology of prejudice.* Saddle River, NJ: Pearson Education, 2002.

Jowell, Roger and Prescott-Clark, Patricia. "Racial Discrimination and White-Collar Workers in Britain." *Race,* November 1970, *11*(4), pp. 397–417.

Lundberg, Shelly J. and Starz, Richard. "Private Discrimination and Social Intervention in Competitive Labor Market." *American Economic Review,* June 1983, *73*(3), pp. 340–47.

McIntyre, Shelby J.; Moberg, Dennis J. and Posner, Barry Z. "Discrimination in Recruitment: An Empirical Analysis: Comment." *Industrial and Labor Relations Review,* July 1980, *33*(4), pp. 543–47.

Mullainathan, Sendhil. "Thinking Through Categories." Mimeo, Massachusetts Institute of Technology, 2003.

Neumark, David. "Sex Discrimination in Restaurant Hiring: An Audit Study." *Quarterly Journal of Economics,* August 1996, *111*(3), pp. 915–42.

Newman, Jerry M. "Discrimination in Recruitment: An Empirical Analysis." *Industrial and Labor Relations Review,* October 1978, *32*(1), pp. 15–23.

Nisbett, Richard E. and Cohen, Dov. *The culture of honor: The psychology of violence in the South.* Boulder, CO: Westview Press, 1996.

Phelps, Edmund S. "The Statistical Theory of Racism and Sexism." *American Economic Review,* September 1972, *62*(4), pp. 659–61.

Raphael, Steven; Stoll, Michael A. and Holzer, Harry J. "Are Suburban Firms More Likely to Discriminate against African Americans?" *Journal of Urban Economics,* November 2000, *48*(3), pp. 485–508.

Riach, Peter A. and Rich, Judity. "Testing for Racial Discrimination in the Labour Market." *Cambridge Journal of Economics,* September 1991, *15*(3), pp. 239–56.

Turner, Margery A.; Fix, Michael and Struyk, Raymond J. *Opportunities denied, opportunities diminished: Racial discrimination in hiring.* Washington, DC: Urban Institute Press, 1991.

Weichselbaumer, Doris. "Sexual Orientation Discrimination in Hiring." *Labour Economics,* December 2003, *10*(6), pp. 629–42.

_____. "Is it Sex or Personality? The Impact of Sex-Stereotypes on Discrimination in Applicant Selection." *Eastern Economic Journal,* Spring 2004, *30*(2), pp. 159–86.

E-lit-006: Meta-analysis of field experiments shows no change in racial discrimination in hiring over time

Lincoln Quillian, Devah Pager, Ole Hexel and Arnfinn H. Midtboen | Proceedings of the National Academy of Sciences 114(41): 10870-10875 | 2017

https://europepmc.org/articles/PMC5642692?pdf=render

About this source

The paper answers two distinct questions and the section's hypothesis needs both. First, the level: pooling all U.S. hiring field experiments run since 1989, whites receive 36 percent more callbacks than otherwise identical African-American applicants (95 percent CI 25 to 47 percent) and 24 percent more than Latinos (CI 15 to 33 percent). That pooled 36 percent is the meta-analytic center of gravity of the literature, and it sits well below the roughly 50 percent white advantage in Bertrand and Mullainathan - the canonical study is toward the high end of the distribution, not the middle of it. Second, the trend: there is none. The line of best fit for African Americans since 1990 is close to flat and slopes slightly upward; the 95 percent confidence interval on the annual slope of the log discrimination ratio is -0.007 to 0.015, so the data rule out anything faster than a 0.7 percent annual decline while the point estimate is no decline at all. The result survives substituting job offers for callbacks (-0.008), dropping applicants with criminal records or disabilities, using resume audits only (-0.002), and adding controls for applicant gender and education, study method, local unemployment, region and occupation, with coefficients across specifications running from +0.001 to +0.013 per year. Latinos are the one place with movement: the point estimate falls from a 1.30 ratio in 1990 to 1.15 in 2010, though at P = 0.099 and on only nine estimates. Publication bias is addressed by including 11 unpublished experiments, whose inclusion barely changes the estimates, and by showing that studies for which race was the incidental rather than the focal variable find no less discrimination.

Bound as the original PDF. Retrieved 2026-09-10. sha256 ddfdcfac783530e3...

Bound into an Economic Briefing · by Claude, may contain mistakes

Full document begins on the next page.

E-lit-006: https://europepmc.org/articles/PMC5642692?pdf=render

Meta-analysis of field experiments shows no change in racial discrimination in hiring over time

Lincoln Quillian[a,b,1], Devah Pager[c,d], Ole Hexel[a,e], and Arnfinn H. Midtbøen[f]

[a]Department of Sociology, Northwestern University, Evanston, IL 60208; [b]Institute for Policy Research, Northwestern University, Evanston IL 60208; [c]Department of Sociology, Harvard University, Cambridge, MA 02138; [d]Kennedy School of Government, Harvard University, Cambridge MA 02138; [e]Sciences Po, Observatoire Sociologique du Changement (OSC), CNRS, 75007 Paris, France; and [f]Institute for Social Research, N-0208 Oslo, Norway

Edited by Douglas S. Massey, Princeton University, Princeton, NJ, and approved August 8, 2017 (received for review April 14, 2017)

This study investigates change over time in the level of hiring discrimination in US labor markets. We perform a meta-analysis of every available field experiment of hiring discrimination against African Americans or Latinos (n = 28). Together, these studies represent 55,842 applications submitted for 26,326 positions. We focus on trends since 1989 (n = 24 studies), when field experiments became more common and improved methodologically. Since 1989, whites receive on average 36% more callbacks than African Americans, and 24% more callbacks than Latinos. We observe no change in the level of hiring discrimination against African Americans over the past 25 years, although we find modest evidence of a decline in discrimination against Latinos. Accounting for applicant education, applicant gender, study method, occupational groups, and local labor market conditions does little to alter this result. Contrary to claims of declining discrimination in American society, our estimates suggest that levels of discrimination remain largely unchanged, at least at the point of hire.

discrimination | labor markets | field experiments | race | ethnicity

The American racial landscape has changed in fundamental ways since the Civil Rights Movement of the 1960s. During that time, sweeping legal and social reforms reduced the barriers facing African Americans in many important domains (1, 2). A rising African American middle class and a growing acceptance of the principles of inclusion led some to conclude that racial discrimination had declined to the point that it was no longer a primary determinant of life chances for African Americans and Latinos (2, 3).

Supporting this perspective, a variety of indicators pointed toward a reduction of discriminatory treatment. Surveys indicated that whites increasingly endorsed the principle of equal treatment regardless of race (4). Rates of high school graduation for whites and African Americans converged substantially, and the black–white test score gap declined (5, 6). Large companies increasingly recognized diversity as a goal and revamped their hiring to curtail practices that disadvantaged minority applicants (7). With the election of the country's first African-American president in 2008, many concluded that the country had finally moved beyond its troubled racial past (8).

Despite clear signs of racial progress, however, on several key dimensions racial inequality persists and has even increased. For example, racial gaps in unemployment have shown little change since 1980 (9, 10), and the black–white gap in labor force participation rates among young men widened during this time (11). Recently, the Black Lives Matter movement shone a spotlight on the ongoing struggles with racism and discrimination experienced by people of color in interactions with law enforcement. The election of Donald J. Trump as the 45th President of the United States with the support of antiimmigrant and white nationalist groups highlighted the persistence of racial resentment (12).

In light of persistent racial gaps in key social and economic indicators, some scholars have challenged prevailing assumptions about waning discrimination. Indeed, while expressions of explicit prejudice have declined precipitously over time, measures of stereotypes and implicit bias appear to have changed little over the past few decades (13–15). In this view, far from disappearing, racial bias has taken on new forms, becoming more contingent, subtle, and covert (15–18).

What can we reliably say about trends in discrimination over time? Has the role of race appreciably diminished across the board, or are there important domains in which little racial progress has been achieved? Answers to these questions are important for understanding the sources of persistent racial inequality.

In this study, we examine trends in racial and ethnic discrimination in American labor markets based on a meta-analysis of every available field experiment of hiring discrimination (with fieldwork dates through December 2015). Meta-analysis is a body of formal methods to synthesize data from a population of existing studies. Field experiments of hiring discrimination are experimental studies in which fictionalized matched candidates from different racial or ethnic groups apply for jobs. These studies include both resume audits, in which fictionalized resumes with distinct racial names are submitted online or by mail (e.g., ref. 19), and in-person audits, in which racially dissimilar but otherwise matched pairs of trained testers apply for jobs (e.g., ref. 20).

The field experimental method is a design with high causal (internal) validity because it benefits from aspects of experimental design. The experimenter carefully manages the application process, which provides control over many potential confounding variables. The exact basis of causal inference across the two main forms of field experiment, resume and in-person audits, is somewhat different. In the typical resume audit, clues indicating race (such as a racially identifiable name) are randomly assigned to otherwise similar resumes, allowing for treatment and control groups to be equated through randomization. In in-person audits, matched

Significance

Many scholars have argued that discrimination in American society has decreased over time, while others point to persisting race and ethnic gaps and subtle forms of prejudice. The question has remained unsettled due to the indirect methods often used to assess levels of discrimination. We assess trends in hiring discrimination against African Americans and Latinos over time by analyzing callback rates from all available field experiments of hiring, capitalizing on the direct measure of discrimination and strong causal validity of these studies. We find no change in the levels of discrimination against African Americans since 1989, although we do find some indication of declining discrimination against Latinos. The results document a striking persistence of racial discrimination in US labor markets.

Author contributions: L.Q. designed research; L.Q., O.H., and A.H.M. performed research; L.Q. and O.H. analyzed data; and L.Q., D.P., O.H., and A.H.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

See Commentary on page 10815.

[1]To whom correspondence should be addressed. Email: l-quillian@northwestern.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1706255114/-/DCSupplemental.

pairs of trained testers who differ on the basis of race but are otherwise similar apply for jobs; the between-race contrast is grounded in matching pairs of applicants to make them as similar as possible in all employment-relevant characteristics except race. Both resume and in-person audit methods provide a strong basis from which to draw conclusions about hiring discrimination, particularly relative to the nonexperimental methods widely used in the literature, including by all prior studies of discrimination trends over time (ref. 21 and *SI Appendix*, section 1).

We use meta-analytic techniques to investigate change in hiring discrimination over time based on all existing US field experimental studies of labor market discrimination. Our procedure follows three basic stages: First, we identified all existing studies, published or unpublished, that use a field experimental method and that provide contrasts in hiring-related outcomes between equally qualified candidates from different racial or ethnic groups. Second, we coded key characteristics of the studies into a database for our analysis based on a coding rubric. This produced 24 studies containing 30 estimates of discrimination against African Americans and Latinos since 1989, together representing 54,318 applications submitted for 25,517 positions. Finally, we performed a random-effects meta-regression to identify trends over time.

We assess discrimination for each study using the ratio of the proportion of applications that received "callbacks"—or invitations to interview—by white applicants relative to African-American or Latino applicants. We calculated the proportions based on counts of the number of callbacks received by each group (white/African American/Latino) within each study. This discrimination ratio measured at the study level is the outcome in our meta-regression. Other methods of calculating hiring disparities between groups produced substantively similar results (*SI Appendix*, section 8).

We analyze the relationship of discrimination ratios to years in which the data were gathered to provide an estimate of the trend in discrimination. Specifically, we regress the log of the discrimination ratio on year of survey, with controls for key characteristics of the studies, using meta-regression. Meta-regression is a procedure similar to standard regression, except covariates are measured at the level of the study rather than the level of the individual, and the outcome is an effect from the study of interest (in our case, the outcome is the estimate of discrimination against African Americans or Latinos). *Methods and Materials* discusses further methodological and modeling details.

Results

To explore trends over time, we estimate a series of meta-regressions. We take the natural log of the discrimination ratio (our outcome variable) to account for skew. In the simplest meta-regression models, the only covariate is the time trend. In later models, we include a more extensive set of predictors to control for other factors that might confound the time trend. To capture sources of variability not covered by the covariates, we use a random effects specification (22). Random effects incorporate a variance component capturing variation in outcomes across studies that are due to unobserved study-level factors (*Methods and Materials*).

Our core analysis focuses on studies that conducted their fieldwork from 1989 to 2015, allowing us to observe trends in discrimination over the past 25 years. For some supplementary analyses, we also add four field experiments conducted before 1989, although these studies use less standardized methodologies. On average, white applicants receive 36% more callbacks than equally qualified African Americans (95% confidence interval of 25–47% more), based on random-effects meta-analysis of data since 1989, representing a substantial degree of direct discrimination. White applicants receive on average 24% more callbacks than Latinos (95% confidence interval of 15–33% more). For more detailed results, see *SI Appendix*, section 2 and Figs. S1 and S2.

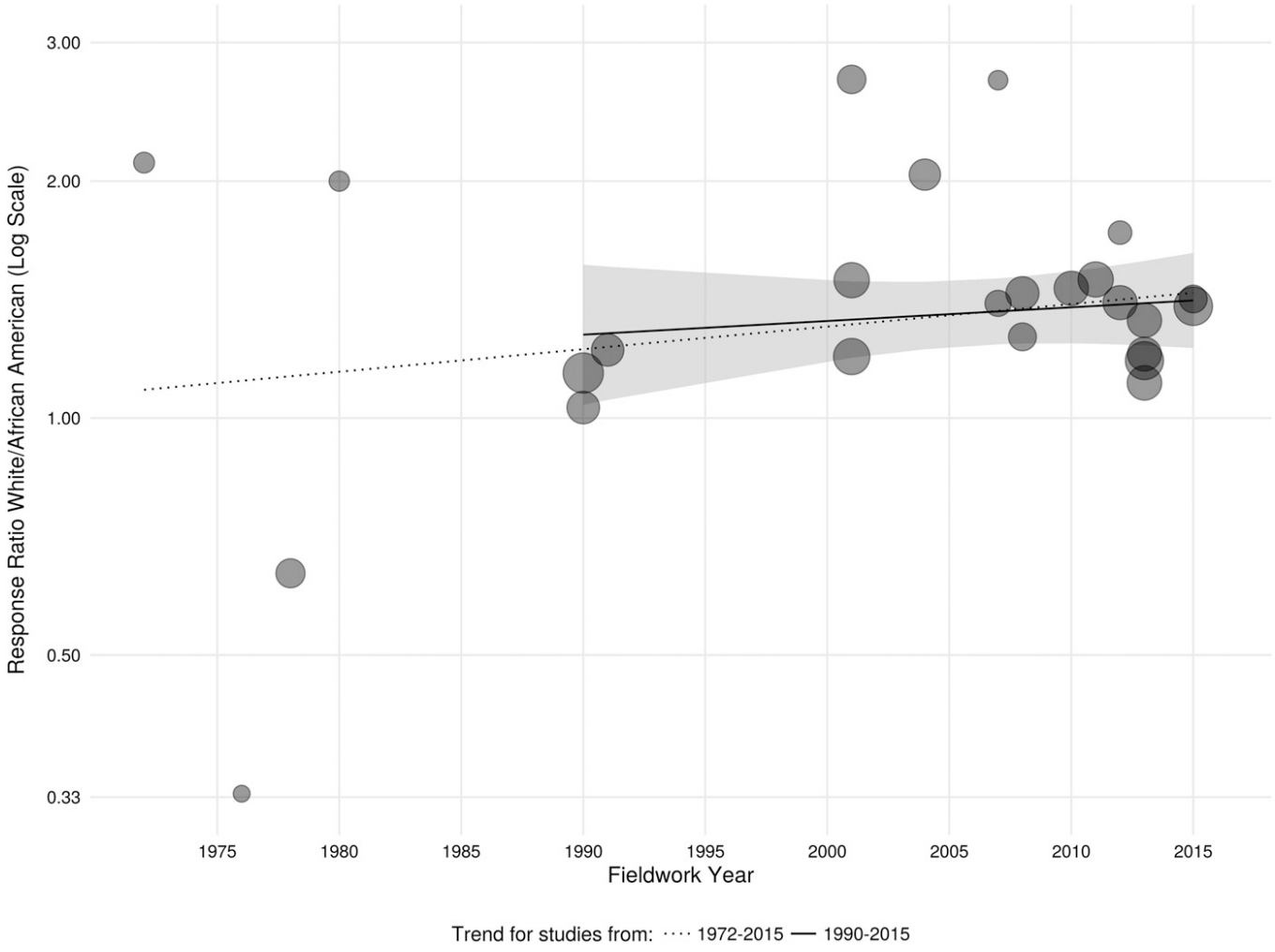
Do we find evidence of change over time in rates of hiring discrimination? With respect to African Americans, the answer is no. Fig. 1 plots estimates of discrimination by year, with linear trends of best fit and 95% confidence regions (detailed estimates are in *SI Appendix*, section 3 and Table S3; in Fig. 1, we exponentiate predictions to present predicted values as discrimination ratios rather than less interpretable log discrimination ratios). The solid line captures the trend since 1990. The dashed line extends this time trend back to 1972, adding four resume audits conducted from 1972 to 1980. The size of the symbol is proportional to the weight it is given in the meta-analysis. The line of best fit for studies since 1990 is close to flat, sloping slightly upward, suggesting no change in the rate of discrimination over the past 25 years. The longer time series includes studies that use a more heterogeneous set of procedures (*Methods and Materials*), but even here we see no clear change over time in the level of hiring discrimination against African Americans.

Is there sufficient power based on 21 studies to conclude that discrimination against African Americans did not decline? The confidence interval of the annual change provides a way to answer this question. The 95% confidence interval of the slope 1989–2015 is −0.007 to 0.015. [This is the confidence interval of the slope of "year" (our time trend variable) with the log discrimination ratio outcome. The regression is shown in *SI Appendix*, Table S3.] The lower end of this interval indicates a decline in the discrimination ratio of 0.7% per year. If we take this number as the smallest slope consistent with the data based on the confidence interval, this suggests only a slight decline in discrimination each year. We conclude that this evidence rules out all but a slow decline in discrimination—with the most likely estimate being the point estimate, which indicates no decline in discrimination at all.

Fig. 2 presents the trend for Latinos (as with Fig. 1, model predictions have been exponentiated to allow interpretation as discrimination ratios rather than log ratios). Here, we see the line slopes downward, indicating a possible decline in discrimination, although this trend is outside of conventional levels of significance ($P = 0.099$). The point estimate suggests a decline from whites receiving 30% more callbacks than Latinos in 1990 to 15% more callbacks in 2010 (1.30 vs. 1.15). Because of the small number of Latino field experiments ($n = 9$), there is high uncertainty in characterizing this trend. (Using the difference in proportions or the odds ratio as outcomes, rather than the discrimination ratio, results in downward slopes in discrimination against Latinos over time that are statistically significant at the $P < 0.05$ level; see *SI Appendix*, section 8 and Table S9. However, sensitivity checks that modified the outcome sample counts slightly result in nonsignificant year coefficients of the difference in proportion or odds ratio, see *SI Appendix*, section 4 and Table S5).

Is it possible that key aspects of study design changed over time, influencing our estimates of changes in discrimination? To consider this question, we estimate a meta-regression model of discrimination rates as a function of a time trend plus other study characteristics. We discuss only models for African Americans, because the number of studies with Latinos ($n = 9$) is too small to produce reasonable precise estimates in a meta-regression model with multiple covariates.

Fig. 3 graphs estimates of change over time when the outcome discrimination ratio is modified and when controls are added. Full coefficients of the models are shown in *SI Appendix*, Table S4, with additional discussion in *SI Appendix*, section 4. The coefficients can be interpreted as the one-year percentage change in the discrimination ratio. [Because the outcome is logged, and $\exp(b) \approx 1 + b$ for $b < 0.1$, coefficients with values less than about 0.1 can be multiplied by 100 to closely approximate percentage changes with a one-unit change in x.] The first coefficient graphed shows the annual percentage change in discrimination from 1990 to 2015, corresponding to solid line in Fig. 1. The second shows the annual change for the longer time period 1972–2015, corresponding to the dashed line in Fig. 1.



Note: Size of plotting symbols proportional to meta-regression weights. Shaded region gives 95% confidence interval.

Fig. 1. No reduction in hiring discrimination facing African Americans over time.

The next few models alter the dependent variable to see if this changes our results (using our base sample of 1990–2015). In one modification, we use "job offer" in place of callback as the outcome for studies for which the job offer outcome is available ($n = 3$), retaining callbacks as the outcome for studies in which the measure of job offer is not available. This makes the outcome variable less uniform across studies, although closer to the outcome of greatest substantive interest, getting a job. With this modification, the trend line for African Americans slants more downward, but is still close to zero (−0.008) and statistically nonsignificant. A second modification eliminates applicant profiles that included either a fictitious criminal background ($n = 7$) or a disability ($n = 1$). This limits the applicant profiles to those with more mainstream job backgrounds and credentials. The modified results show the trend line slanting slightly more upward, providing less evidence to support a downward trend than the results including a more heterogeneous set of applicant characteristics. A third modification uses only resume audit studies, discarding in-person audits. This results in an almost perfectly flat line (−0.002).

The next estimates are based on models that add controls for applicant attributes, region and area unemployment rates, and occupational categories to the baseline time-trend model. The "Applicant Attributes" model introduces covariates representing applicant characteristics (e.g., gender, education) and study design (resume or in-person audits). The "UE & Regions" model adds controls for the unemployment rate of the local metropolitan area and dummy variables for region. The "Occupations" model includes controls for occupational categories of blue collar, office-focused, and restaurant occupations. Finally we present the coefficient from a trimmed model in which only the predictors with the largest t ratios from prior models are included. In each case, we see coefficients for the time trend that are close to zero—ranging from an estimated increase of 0.1% per year (0.001) to an increase of 1.3% per year (0.013) —suggesting little change in the level of discrimination facing African Americans over time. Notably, then, we find evidence of stability, not change, in hiring discrimination against African Americans.

Few of the measured covariates in our analysis (*SI Appendix*, Table S4) demonstrate a clear relationship to patterns of discrimination. This likely is in part due to the relatively small overall sample of studies ($n = 21$ for African Americans since 1990), which limits our ability to detect statistical significance. However, even looking at the point estimates we find no large differences in magnitude across categories. This result is consistent with the findings within individual audit studies that suggest relative stability in measured discrimination across job types, applicant gender, and skill levels (e.g., refs. 19 and 20). (We also note that a meta-analysis designed to look specifically at effects of many of

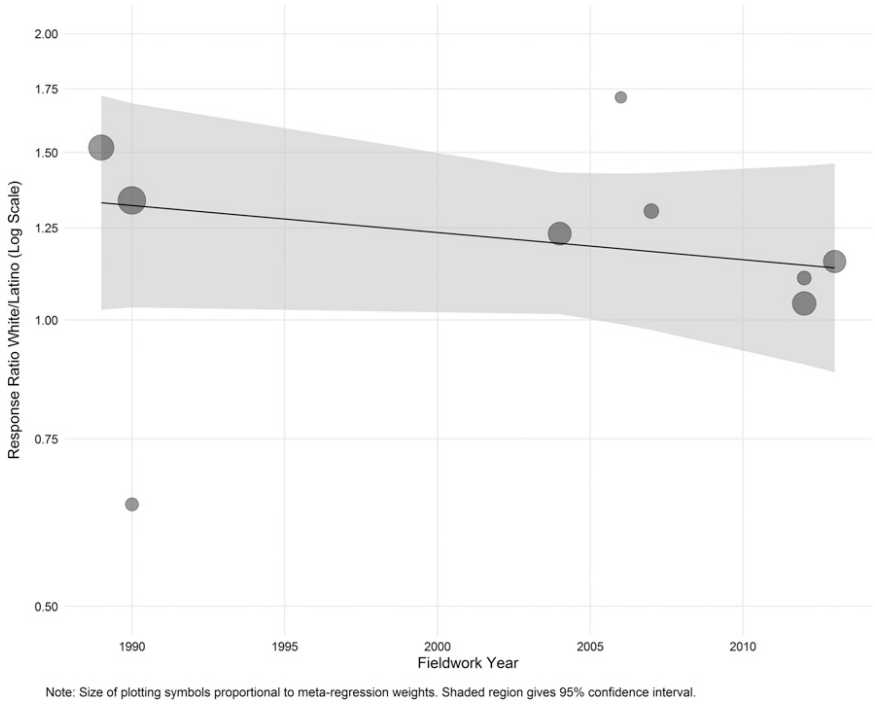


Fig. 2. Modest evidence of a reduction in hiring discrimination facing Latinos over time.

these covariates would use within-study variability—such as contrasting male and female auditors in the same study—which could provide more power to discern effects. Within-study variability cannot be applied to understand change over time since studies are generally conducted over the span of just a few months.) As a final check on the influence of covariates, we tested for time trends among our study-level and individual-level characteristics, finding no evidence of systematic change (*SI Appendix*, section 9 and Table S10). This suggests that covariates are unlikely to influence the observed time trend for discrimination among either the African-American or Latino samples.

In relation to our estimate of changes in discrimination over time, the inclusion of study-level and applicant-level characteristics has little impact. In all models, we see little evidence of a reduction in hiring discrimination against African Americans over time.

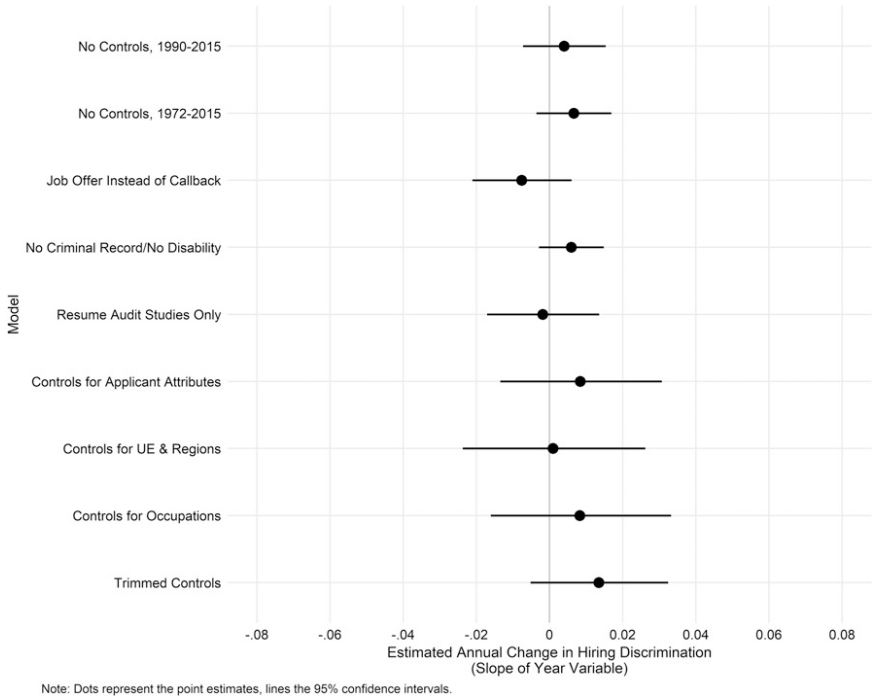


Fig. 3. Alternative estimates of change in hiring discrimination against African-Americans.

A potential concern of any meta-analysis is publication bias. In the present case, publication bias may entail studies that show no discrimination being less likely to be published and, thus, included in our study. We sought to address this issue by seeking out and including all nonpublished field experiments available ($n = 11$). Their inclusion did little to affect our estimates. Finally, in *SI Appendix*, section 5 and Table S7, we show that studies in which racial discrimination was the focus of the analysis (and for which there may be more pressure to demonstrate a positive effect) show no more discrimination than studies in which other characteristics were the main focus (with race included as an secondary or incidental covariate), further reducing concerns over publication bias for our results.

Discussion

Contrary to widespread assumptions about the declining significance of race, the magnitude and consistency of discrimination we observe over time is a sobering counterpoint. We note that our results do not address the possibility that hiring discrimination may have substantially dropped in the 1960s or early 1970s, during the civil rights era when many forms of direct discrimination were outlawed, as some evidence suggests (1). Further, we note that our results pertain only to discrimination at the point of hire, not at later points in the employment relationship such as in wage setting or termination decisions. Social psychological theories would predict hiring to be most vulnerable to the influence of racial bias, given that objective information is limited or unreliable (23–25). Likewise, from an accountability standpoint, discrimination is less easily detected, and therefore less costly to employers, at the point of hire (26). It may be the case, then, that more meaningful reductions in discrimination have taken place at other points in the employment relationship not measured here. What our results point to, however, is that at the initial point of entry—hiring decisions—African Americans remain substantially disadvantaged relative to equally qualified whites, and we see little indication of progress over time.

These findings lead us to temper our optimism regarding racial progress in the United States. At one time it was assumed that the gradual fade-out of prejudiced beliefs, through cohort replacement and cultural change, would drive a steady reduction in discriminatory treatment (27). At least in the case of hiring discrimination against African Americans, this expectation does not appear borne out.

We find some evidence of a decline in discrimination against Latinos since 1989. The small number of audit studies including Latinos limits our ability to include controls and the precision of our estimates—the decline is marginally significant statistically ($P = 0.099$). More evidence is needed to establish the trend in hiring discrimination against Latinos with greater certainty.

Our results point toward the need for strong enforcement of antidiscrimination legislation and provide a rationale for continuing compensatory policies like affirmative action to improve equality of opportunity. Discrimination continues, and we find little evidence in regards to African Americans that it is disappearing or even gradually diminishing. Instead, we find the persistence of discrimination at a distressingly uniform rate.

Materials and Methods

Our procedure follows three basic stages: first, to identify all existing field experiments of hiring discrimination; second, to develop a coding rubric and to code studies to produce a database of their results; and third, to perform a statistical meta-analysis to draw conclusions from the combined results. We discuss each of these steps in turn.

Identifying Relevant Studies. We aimed to include in our meta-analysis all existing studies, published or unpublished, that use a field experimental method and that provide contrasts in hiring-related outcomes between different race and ethnic groups in the United States. This includes both in-person audit studies and resume studies (or correspondence studies). We also required that contrasts of hiring outcomes between race or ethnic groups were made for groups that were on average equivalent in their labor market relevant characteristics, since otherwise discrimination estimates are confounded with the difference in nonracial characteristics.

We used three methods to identify relevant field experiments: searches in bibliographic databases, citation searches, and an email request to corresponding authors of field experiments of race-ethnic discrimination in labor markets and other experts on field experiments and discrimination.

We began with a bibliographic search. Our search covered the following bibliographic databases and working paper repositories: Thomson's Web of Science (Social Science Citation Index), ProQuest Sociological Abstracts, ProQuest Dissertations and Theses, Lexis Nexis, Google Scholar, and NBER working papers. We searched for some combination of "field experiment" or "audit study" or "correspondence study" and sometimes included the term "discrimination," with some variation depending on the search functions of the database. We also searched two French-language indexes, Cairn and Persée, and two international sources, IZA discussion papers, a German working paper archive, and ILO International Migration Papers.

Our second technique for identifying relevant studies relied on citation search. Working from the initial set of studies located through bibliographic search, we examined the bibliographies of all review articles and eligible field studies to find additional field experiments of hiring discrimination.

The last technique used was an email request of authors of existing field experiments of discrimination. From our list of audit studies identified by bibliographic and citation search, we compiled a list of email addresses of authors of existing field experiments of discrimination. To this we added the addresses of authors of literature review articles on field experiments. Our email request asked for citations or copies of field discrimination studies published, unpublished, or ongoing. We also asked that authors refer us to any other researchers who may have recent or ongoing field experiments.

The email requests were conducted in two phases. In the initial wave, 131 apparently valid email addresses were contacted. We received 56 responses. We also sent out a second wave of 68 e-mails which consisted of additional authors identified from the initial wave of surveys and some corrected email addresses. We received 19 responses to this second wave of email surveys.

Overall, our search located 34 studies that were US-based field experiments of hiring, included contrasts between white and nonwhite applicant profiles that were on-average equivalent in their labor-market relevant characteristics (e.g., education, experience level in the labor market). Six studies were excluded for various reasons, as explained in *SI Appendix*, section 6. Our remaining 28 studies yielded 24 estimates of discrimination against African Americans and 9 against Latinos relative to whites.

Coding and Selection of Analysis Period (1989–2015). We coded key characteristics of the studies into a database for our analysis. Coding was based on a coding rubric, which listed each potentially relevant characteristic of the research and included coding instructions. To develop the rubric, we initially read several studies and, based on this, developed an initial coding rubric of factors we thought might influence measured rates of discrimination. The initial rubric was reviewed and updated by all authors of this study for completeness. It was subsequently refined as coding progressed. Each study was coded independently by two raters, with disagreement resolved by the first author. See *SI Appendix*, section 7 for more discussion of coding procedures. A list of coded characteristics for the 1989–2015 studies are shown in the *SI Appendix*, Tables S1 and S2.

Studies have fieldwork periods range from 1972 to 2015 for African Americans and 1989 to 2015 for Latinos. For most analyses in this paper, we focus on the period 1989–2015. We focus on this period because the data are sparse before this period (only four studies before 1989) and because our reading of the early studies indicates key methodological differences among these early studies that may affect their results. Resume audits typically signal race by using race-typed names on resumes, but the pre-1989 studies either indicated race directly on the resume [McIntyre et al. (28) put "Race: BLACK" on the minority resumes and nothing about race on the "white" resumes] or attached photos to resumes (a procedure used by Newman; ref. 29). Excluding the early studies leaves us with 21 estimates of discrimination against African Americans and nine against Latinos from 24 studies (six studies include estimates of discrimination against both African Americans and Latinos).

The Meta-Analysis Model. A meta-analysis aggregates information from across studies to produce an estimate of an effect of interest (30). In this study, our basic measure of discrimination is the discrimination ratio. This is the ratio of the percentage of callbacks for interviews received by white applicants to the percentage of callbacks or interviews received by African Americans or Latinos. Formally, if c^w is the number of callbacks received by whites, and c^m is the number of callbacks received by African Americans or Latinos, and n^w is the number of applications submitted by white applicants, and n^m is the

number of applications submitted by African American or Latino applicants, then the discrimination ratio is $(c^w/n^w)/(c^m/n^m)$. Ratios above 1 indicate whites received more positive responses than African Americans or Latinos, with the amount above 1 multiplied by 100 indicating the percentage higher callbacks for whites relative to the minority group. Because audit studies equate groups on their nonracial characteristics either through matching and assignment of characteristics (in-person audits) or through random assignment (most resume audits), no further within-study controls are required. *SI Appendix*, section 8 discusses potential alternative measures of discrimination using the difference in proportions and the odds ratio, and presents alternative results using these measures. Our basic result—no decline in discrimination against African Americans over time—holds using both of the alternative measures, whereas evidence of a decline in discrimination for Latinos appears somewhat stronger with the difference in proportions or the odds ratio.

The goal of a meta-analysis is to combine information across studies. This requires measuring the information each study contains about discrimination against a group. The information each study provides is inversely proportional to the square of the SE of the discrimination ratio. We calculate the SE of the ratio from counts reported in each study, accounting for audit pairs in the design when possible. In cases where information on paired outcomes is available from the study (counts of pairs in which both the white and the nonwhite tester receive a callback, white yes nonwhite no, white no nonwhite yes, neither get a callback), we calculated SEs of discrimination ratios accounting for the pairing (see *SI Appendix*, section 9 for details and formulas). For studies that are not paired between whites or nonwhites or where paired outcomes are not reported, we use formulas for the SE for unpaired groups. This formula will slightly overestimate the SE of the effect for studies that are paired but we treat as unpaired due to lack of information about the outcomes at the pair level, underweighting these studies a bit in computing the overall effect, and slightly inflating the overall cross-study SE.

Of course field experiments vary in their characteristics, such as the geographic area they cover, the exact job sectors covered, and details of their methodology. To account for this variability in understanding the time trend, we use two procedures. First, we include controls, discussed further below, for many study characteristics. Second, to capture sources of variability not covered by the covariates, we use a random effects specification (22). Random effects incorporate a variance component capturing variation in outcomes across studies that are due to unobserved study-level factors. Random effects are recommended whenever there is reason to believe that the effect in question is likely to vary as a function of design features of the study, rather than representing a single underlying effect that is constant over the whole population. This is surely the case in our analysis, as we expect that the level of racial discrimination may depend on the year of the study, the situation the study considers (e.g., the occupational categories), the skill level of the applicants, and so on. The random effect increases the SEs of estimates to correctly account for variabilities among studies in drawing inferences about overall trend.

More formally, random-effects meta-analysis allows the true effects of race on the callback rate in each situation estimated by each study, θ_i, to vary between studies by assuming that they have a normal distribution around a mean effect, θ. If y_i is the discrimination ratio in the ith study, then the meta-analysis model is as follows:

$$\ln(y_i) = \theta + u_i + e_i, \text{where } u_i \sim N(0, \tau^2) \text{and } e_i \sim N(0, \sigma_i^2).$$

Here, τ^2 is the between-study variance, estimated from between-study variance as part of the meta-analysis model, while σ_i^2 is the variance of the log response ratio in the ith study, estimated from study counts as described above. Following standard practice in the meta-analysis literature, we log the response ratio to reduce the asymmetry of the ratio.

Meta-regression allows that the rate of discrimination is a function of a vector of k characteristics of the studies and effects, x, plus (in the random effects specification) residual study-level heterogeneity (between study variance not explained by the covariates). The model assumes the study-level heterogeneity follows a normal distribution around the linear predictor:

$$\ln(y_i) = x_i\beta + u_i + e_i, \text{where } u_i \sim N(0, \tau^2) \text{and } e_i \sim N(0, \sigma_i^2),$$

where β is a $k \times 1$ vector of coefficients (including a constant), and x_i is a $1 \times k$ vector of covariate values in study i (including a 1 for a constant). Estimation is by restricted maximum likelihood. For details, see *SI Appendix*, section 9.

To explore trends over time, we include covariates for the year of fieldwork of the study. In the simplest models, the only covariate is this time trend. In later models, we include a more extensive set of predictors to control for other factors that might confound the time trend. These additional controls include resume audit vs. field audit as the study method, gender and education level of the fictitious applicants, occupations tested, unemployment rates at the field sites used for testing, criminal background of some fictitious applicants, and region of the country. For discussions of why these controls were selected, see *SI Appendix*, section 4 (see *SI Appendix*, Tables S1 and S2 for descriptive statistics on the controls; for a discussion of trends in covariates, see *SI Appendix*, section 10).

ACKNOWLEDGMENTS. We thank Anthony Heath, Fenella Fleischmann, Matthew Salganik, Frank Dobbin, András Tilcsik, Donald Green, David Neumark, Hedwig Lee, two anonymous PNAS reviewers, and the editor for comments; Larry Hedges for methodological advice; and Jim Cheng Chen and Joshua Aaron Klingenstein for excellent research assistance. We have received financial support for this project from the Russell Sage Foundation and the Institute for Policy Research at Northwestern University.

1. Donohue J-J, Heckman J (1991) Continuous versus episodic change: The impact of civil rights policy on the economic status of blacks. *J Econ Lit* 29:1603–1643.
2. Wilson W-J (1978) *The Declining Significance of Race: Blacks and Changing American Institutions* (Univ of Chicago Press, Chicago).
3. Heckman J (1998) Detecting discrimination. *J Econ Perspect* 12:101–116.
4. Schuman H, Steeh C, Bobo L, Krysan M, eds (1998) *Racial Attitudes in America: Trends and Interpretations* (Harvard Univ Press, Cambridge, MA), Rev ed.
5. Jencks C, Phillips M (2011) *The Black-White Test Score Gap* (Brookings Institution Press, Washington, DC), p 542.
6. Reardon S-F, Fahle E (2017) Education. State of the union: The poverty and inequality report (Stanford Center on Poverty and Inequality, Palo Alto, CA), special issue, *Pathways* Magazine.
7. Dobbin F (2009) *Inventing Equal Opportunity* (Princeton Univ Press, Princeton, NJ), p 321.
8. Tesler M, Sears D-O (2010) *Obama's Race: The 2008 Election and the Dream of a Post-Racial America* (Univ of Chicago Press, Chicago).
9. Austin A-A (2013) The unfinished march: An overview. (Economic Policy Institute). Available at www.epi.org/publication/unfinished-march-overview/. Accessed May 25, 2016.
10. Cancio A-S, Evans D, Maume D-J (1996) Reconsidering the declining significance of race: Racial differences in early career wages. *Am Sociol Rev* 61:541–556.
11. Holzer H-J, Offner P, Sorensen E (2005) Declining employment among young black less-educated men: The role of incarceration and child support. *J Policy Anal Manage* 24:329–350.
12. Hajnal Z, Abrajano M (2016) Trump's all too familiar strategy and its future in the GOP. *Forum* 14:295–309.
13. Devine P-G, Elliot A-J (1995) Are racial stereotypes really fading? The Princeton trilogy revisited. *Pers Soc Psychol Bull* 21:1139–1150.
14. Bobo L-D, Charles C-Z, Krysan M, Simmons A-D (2012) The real record on racial attitudes. *Social Trends in American Life: Findings from the General Social Survey Since 1972*, ed Marsden P-V (Princeton Univ Press, Princeton), pp 38–83.
15. Dovidio J-F, Gaertner S-L (2010) Intergroup bias. *Handbook of Social Psychology*, eds Fiske S-T, Gilbert D-T, Lindzey G, Jongsma A-E (Wiley, Hoboken, NJ), 5th Ed.
16. Kinder D-R, Sears D-O (1981) Prejudice and politics: Symbolic racism versus racial threats to the good life. *J Pers Soc Psychol* 40:414–431.
17. McConahay J-B (1983) Modern racism and modern discrimination the effects of race, racial attitudes, and context on simulated hiring decisions. *Pers Soc Psychol Bull* 9: 551–558.
18. Bonilla-Silva E (2006) *Racism Without Racists: Color-Blind Racism and the Persistence of Racial Inequality in America* (Rowman & Littlefield, Lanham, MD), 2nd Ed.
19. Bertrand M, Mullainathan S (2004) Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *Am Econ Rev* 94:991–1013.
20. Pager D, Western B, Bonikowski B (2009) Discrimination in a low-wage labor market: A field experiment. *Am Sociol Rev* 74:777–799.
21. National Research Council (2004) *Measuring Racial Discrimination*, eds Blank RM, Dabady M, Citro CF (National Academies Press, Washington, DC).
22. Raudenbush S-W (2009) Analyzing effects sizes: Random effects coding. *The Handbook of Research Synthesis and Meta-Analysis*, eds Cooper HM, Hedges LV, Valentine JC (Russell Sage Foundation, New York), 2nd Ed, pp 295–316.
23. Oettinger G-S (1996) Statistical discrimination and the early career evolution of the black-white wage gap. *J Labor Econ* 14:52–78.
24. Fiske S (1998) Stereotyping, prejudice, and discrimination. *The Handbook of Social Psychology*, eds Gilbert D, Fiske S, Lindzey G (Wiley, New York), 4th Ed, pp 357–411.
25. Altonji J-G, Pierret C-R (2001) Employer learning and statistical discrimination. *Q J Econ* 116:313–350.
26. Wessel D (September 11, 2003) Fear of bias suits may be affecting hiring decisions. *Wall Street Journal*, section A, p 2.
27. Firebaugh G, Davis K-E (1988) Trends in antiblack prejudice, 1972–1984: Region and cohort effects. *Am J Sociol* 94:251–272.
28. McIntyre S, Moberg D-J, Posner B-Z (1980) Preferential treatment in preselection decisions according to sex and race. *Acad Manage J* 23:738–749.
29. Newman J-M (1978) Discrimination in recruitment: An empirical analysis. *Ind Labor Relat Rev* 32:15–23.
30. Borenstein M, Hedges L-V, Higgins J-P-T, Rothstein H-R (2009) *Introduction to Meta-Analysis* (John Wiley & Sons, Chichester, UK), p 421.

E-lit-007: Systemic Discrimination Among Large U.S. Employers

Patrick M. Kline, Evan K. Rose and Christopher R. Walters | Quarterly Journal of Economics 137(4): 1963-2036 (NBER Working Paper 29053) | 2022

https://www.nber.org/system/files/working_papers/w29053/w29053.pdf

About this source

Applications carrying distinctively Black names were contacted 2.1 percentage points less often than otherwise identical applications with distinctively white names (25.1 percent versus 23.0 percent), a difference significant at p < 10^-32. That is a 9 percent proportional penalty - far below the roughly 36 percent white advantage Quillian et al. (2017) report for recent U.S. correspondence experiments, and far below Bertrand and Mullainathan's 50 percent. The authors are explicit that this is arithmetic, not a change in employer behavior in levels: the level gap of about 2 percentage points sits squarely alongside Nunley et al. (2015) at 2.6 points and Agan and Starr (2018) at 2.4 points, and the proportional gap is smaller only because these very large employers contacted 24 percent of all applicants, nearly three times Bertrand and Mullainathan's rate. They offer one substantive reason it could also be genuinely smaller - very large firms may discriminate less, as a Canadian experiment found. The penalty is stable rather than fading: white applicants were contacted more often in 12 of 13 study months, and constancy across the five waves cannot be rejected. The average gap conceals enormous firm heterogeneity - a between-firm standard deviation of 1.9 percentage points, almost as large as the mean - with the top quintile of firms responsible for 46 percent of all lost contacts, a Gini of 0.40, and 23 individual companies identifiable as discriminators while controlling the false discovery rate at 5 percent. By contrast the average gender gap is zero, but the between-firm standard deviation of 2.7 points shows some firms favor men and others women, and applicants over 40 face a 0.6 point penalty.

Bound as the original PDF. Retrieved 2026-09-10. sha256 015e29ffa702a9f0...

Bound into an Economic Briefing · by Claude, may contain mistakes

Full document begins on the next page.

NBER WORKING PAPER SERIES

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, Revised May 2022

We thank Sendhil Mullainathan, Sapna Raj, and Jenny Yang for helpful conversations, our discussants Peter Hull and Peter Q. Blair, and seminar participants at UC Berkeley, Chicago Booth, the University of Chicago HCEO Working Group, the University of Wisconsin, CEPR, IZA, the NBER Labor Studies meetings, 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, Elena Marchetti-Bowick, 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.

© 2021 by Patrick M. Kline, Evan K. Rose, and Christopher R. Walters. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Systemic Discrimination Among Large U.S. Employers
Patrick M. Kline, Evan K. Rose, and Christopher R. Walters
NBER Working Paper No. 29053
July 2021, Revised May 2022
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 discrimination against distinctively Black names 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

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

Evan K. Rose
Becker Friedman Institute
University of Chicago
Chicago, IL 60637
and NBER
ekrose@uchicago.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. Yet a large social science literature analyzing resume correspondence experiments finds 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). In a reanalysis of several correspondence experiments, Kline and Walters (2021) find that discriminatory biases vary tremendously across job vacancies. Less is known, however, about the extent to which discriminatory jobs are concentrated in particular companies. 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. First, 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 employers. Second, if the most biased firms also tend to offer the highest wages, the contribution of discrimination to observed disparities will tend to be amplified (Card, Cardoso and Kline, 2016; Gerard et al., 2021). Third, if only a few firms discriminate, and do so heavily, it may be possible for government regulators to target these companies for audits and investigations. For instance, the Office of Federal Contract Compliance (OFCCP) annually audits thousands of federal contractors for compliance with equal employment laws (Maxwell et al., 2013). Likewise, the U.S. Equal Employment Opportunity Commission (EEOC) routinely launches investigations into whether particular companies have engaged in "systemic discrimination," a term they define 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*).

This paper reports the results of a massive nationwide correspondence experiment designed to measure patterns of discrimination by large U.S. companies. The twin goals of our analysis are to quantify the extent to which discriminatory patterns differ across firms and to assess the feasibility of using experimental evidence to target firms likely to be engaged in discrimination. To facilitate these goals, our experiment was designed to repeatedly elicit signals of bias from specific companies. Unlike traditional audit studies that passively sample jobs from newspapers or job boards (e.g., Bertrand and Mullainathan, 2004), we prospectively applied to entry level job vacancies hosted on the web portals of 108 Fortune 500 firms. For each company, we sampled up to 125 entry-level

jobs in distinct U.S. counties. 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 consistent organization-wide patterns.

Following a large social science literature (Bertrand and Duflo, 2017; Baert, 2018), our 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. Over 83,000 job applications were sent in total, providing uniquely precise signals of employer conduct.

Overall, 24% of the applications we sent were contacted by employers within 30 days. This contact rate is nearly three times greater than what Bertrand and Mullainathan (2004) found in their seminal experiment, suggesting that our fictitious applicants were viewed as plausible job candidates by employers. 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. Past work has typically found larger proportional effects, which may be attributable both to less biased behavior among the extremely large employers we study and the high overall contact rates yielded by our experiment.

A key finding of our analysis is that patterns of discrimination against Black names vary substantially across employers. 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. Despite this wide variability, we cannot reject the null hypothesis that all 108 firms in our experiment weakly favor white names. An application of Efron's (2016) empirical Bayes (EB) deconvolution estimator reveals that while most firms exhibit mild discrimination against Black applicants a few exhibit very large biases. We estimate that the top quintile of discriminating firms are responsible for nearly half of the lost contacts to Black applicants in our experiment. The Gini coefficient of employer contact gaps is estimated to be approximately 0.4, suggesting that discrimination against Black names is roughly as concentrated among firms in our experiment as income is among U.S. households.

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 large U.S. employers exhibit corresponding cross-company patterns of heterogeneity in their average gender contact gaps. Like racial discrimination, gender discrimination is highly concentrated in particular firms, with the top quintile of discriminating firms responsible for nearly 60% of contacts lost to gender discrimination and a Gini concentration coefficient of roughly 0.5.

While our main focus is on race and gender, we also assess the extent of discrimination on several other dimensions. A modest contact penalty of 0.6 percentage points is found 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.

Surprisingly, geographic variation in race, gender, and age discrimination is relatively muted. We cannot reject the null hypothesis that mean contact gaps for gender and age are equal across all 50 states, and find only marginally significant evidence against this null for racial contact gaps. In contrast, 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, suggesting that screening algorithms are unlikely to drive the firm differences we measure.

Consistent with classic models of customer discrimination, both racial and gender contact gaps are estimated to be larger in sectors intensive in jobs requiring social interaction. In line with the predictions of Becker (1957), racial contact gaps are smaller at more profitable firms. Racial contact gaps also tend to be smaller among federal contractors, which is consistent with Miller (2017)'s finding that contracting with the federal government yields sustained increases in Black employment. 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 organization wide biases.

The finding of significant employer heterogeneity in discriminatory conduct motivates an investigation of which particular organizations are likely violating the Civil Rights Act. As a first approach to characterizing detection possibilities, we form EB posterior mean estimates of the contact gap at each individual firm. Firms with posterior mean contact gaps in the top quartile of the distribution are estimated to account for roughly half of the contacts lost to racial discrimination. Discrimination is disproportionately clustered in customer facing sectors including the auto services and sales sector and certain forms

of retail. We find large posterior mean contact gaps favoring women at apparel stores and slightly less pronounced gaps 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 result implies that at least 22 of these 23 firms should be expected to exhibit 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 detect any firms that discriminate against women when 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 conclude are likely 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, indicating a nationwide pattern of discrimination against Black names.

We conclude with an economic analysis of optimal auditing strategies meant to mimic the objectives and constraints of regulatory authorities such as the EEOC or OFCCP. Building on the framework introduced in Kline and Walters (2021), 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 23 firms we classify as discriminating against Black names are estimated to account for nearly 40% of lost contacts to Black applicants in our experiment.

Our findings demonstrate that it is possible to target the individual firms responsible

for a substantial share of discrimination against Black names while maintaining a tight limit on the expected number of false positives. The evidence of discriminatory patterns uncovered here can, in principle, be used by organizations such as the EEOC or OFCCP to target audits and investigations more effectively. Alternatively, this information can be shared directly with the firms themselves, or even made public, potentially enabling companies to preemptively reform their practices, perhaps by adopting the recruiting policies of their less discriminatory peers.

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-qualified 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 disparate treatment, however motivated. The correspondence experiment we study was designed to manipulate employer percep-

[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`).

tions 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] Although it is unclear whether the statistical evidence provided in an audit study would, on its own, be sufficient to successfully litigate a Title VII disparate treatment claim, such evidence may be helpful in building a case or in targeting investigations that lead to the discovery of additional evidence that eventually proves decisive.[3] Conversely, correspondence evidence suggesting equal treatment of workers with different characteristics could, in principle, be used by firms to *counter* charges of employment discrimination. However, further evidence would likely be required for such a determination, as audit studies may fail to detect biases that manifest only at later stages of the hiring process or among applicants with qualification levels outside those considered in the study.

While the social science literature has proposed several distinct theories and definitions of systemic discrimination (e.g., Pincus, 1996; Reskin, 2012), our usage of this phrase is motivated by the EEOC's definition of this term as a "pattern or practice" of discrimination (U.S. EEOC, 2006*b*; Kim, 2015). 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.[4] 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 discrimination against protected characteristics 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.[5] Evidence

[2]For discussion of the potential legal ramifications of handling fictitious applications based upon racial perceptions of names, 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). 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]Explicit evidence of intent to discriminate is *not* required to establish a prima facie case for disparate treatment. The EEOC's own guidance states that "discriminatory motive can be inferred from the fact that there were differences in treatment" (International Brotherhood of Teamsters v. United States, 431 U.S. 324 (1977)). In some cases, large statistical disparities alone can also constitute prima facie evidence of intentional discrimination (Hazelwood School Dist. v. United States, 433 U.S. 299 (1977)).

[4]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).

[5]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

of patterns of discrimination by federal contractors is especially pertinent to the OFCCP, which has broad discretion to audit contractors for compliance with executive orders prohibiting employment discrimination and regularly levies fines and, in some cases, even debars contractors when violations are found (Maxwell et al., 2013).

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).[6] 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. A finding that effects of this nature are present across a large set of establishments suggests 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 sufficient geographic variation and routinely posted enough entry-level jobs on an easily-accessible portal to satisfy our sampling criteria. These 108 large firms, 10 of which are subsidiaries of parent companies in the Fortune 500, employed roughly 15 million workers in 2020 according to Compustat and cover a wide array of industries detailed later on in Table 10.

We sampled 125 entry-level job vacancies from each employer, with each vacancy corresponding to an establishment in a different U.S. county. Sampling was organized in

landlords (Fang, Guess and Humphreys, 2019) and employers (Pager, Bonikowski and Western, 2009).

[6]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."

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 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.[7] 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 that of 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% 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.[8] 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 uni-

[7]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.

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

formly 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 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.[9]

[9]The pre-analysis plan is stored in the AEA RCT registry with number AEARCTR-0004739.

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 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 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 reports 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 cat-

egory ($p \geq 0.24$). 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. Of course, differences in employer treatment of distinctively Black and white (or male and female) names may, in part, reflect stereotypes about average productivity differences between these groups. We leave an investigation of this possibility to future work. Note however that the courts—not to mention potential customers, employees, and corporate shareholders—are likely to view claims that an employer discriminates against applicants with Black (or female) names based on productivity grounds as a pretext for illegal discrimination.

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 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 esti-

mates 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 the effects of LGBTQ club membership 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 at the 10% level ($p = 0.065$), 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

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

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, 1988; Aaronson, Barrow and Sander, 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 Testing for firm components

To test formally for the significance of firm-level contact gap variation, we report a 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. To improve power, we instead employ the high-dimensional moment inequality testing procedure of Bai, Santos and Shaikh (2021), which drops firms with strongly positive z-scores.

The first two columns of Table 4 report 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 suggests 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 result strongly suggests the presence of 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 ($p \leq 0.05$). These findings extend and corroborate recent work by Kline and Walters (2021) and Hangartner, Kopp and Siegenthaler (2021) who conclude that gender discrimination varies bi-directionally across jobs in Mexico and Switzerland, respectively.

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. By contrast, the null hypothesis that no firms discriminate against older applicants is rejected at conventional levels ($p = 0.03$).

6.3 Variance component 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. Graphical evidence of the cross-wave stability of race, gender, and age contact gaps is provided in Appendix Figure A6, which plots firm contact gaps in each wave against their leave-wave-out means. These plots also reveal that firm contact gaps for race and gender are not significantly correlated with each other.

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.

6.4 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 0.8 percentage points, less than half the magnitude of the between-firm standard deviation reported in Table 4.

Controlling for firm fixed effects reduces the modest state variation in contact gaps even further. Table 7 uses the leave-out estimator 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 $(0.190/0.267)^2 \times 100 = 51\%$ 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.[10]

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 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.[11] 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

[10]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.

[11]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.

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. In unreported results, we also found no significant difference in contact gaps between firms that required a battery of cognitive and personality tests and those that did not. The platforms themselves therefore do not appear to be an important driver of the between-firm differences we document.

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 nonetheless offer a low-dimensional summary of the sorts of jobs, establishments, and firms where discrimination tends to be more or less severe. Figures 3, 4, and 5 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 established that contact gaps vary substantially across job titles but that this variation is insignificant conditional on firm effects. While this finding suggests that variation in discrimination across job titles is mostly attributable to the identity of the parent firm, studying lower dimensional summaries of job titles may allow detection of more subtle relationships. A large literature (e.g., Deming, 2017; Hurst, Rubinstein and Shimizu, 2021) finds that the task content of work provides a useful summary of changes in the occupational structure of wages and employment. Adopting this approach, Figure 3 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 correlation may reflect employer concerns regarding customer discrimination, the quantitative importance of which has proven difficult to establish decisively (Holzer and Ihlanfeldt, 1998; Leonard, Levine and Giuliano, 2010; Hurst, Rubinstein and Shimizu, 2021). 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 ($p = 0.20$). This finding casts doubt on simplistic versions of the customer discrimination hypothesis where all employers discriminate differentially in customer facing jobs. For gender, the task content variables are marginally significant conditional on firm fixed effects ($p = 0.01$), suggesting that, at a typical large firm, men face discrimination in customer facing jobs while women face discrimination at jobs intensive in manual skills.

Appendix Figure A7 decomposes the relationship between contact gaps and job task content into within and between industry components. Within industry relationships between racial contact gaps and task content are weak and statistically insignificant, indicating that the task content correlations documented in Figure 3 are driven primarily by between industry variation. Contact gaps are especially strongly related to industry average customer interaction scores ($p = 0.001$). In contrast, the relationship between gender contact gaps and task content is strong both within and between industries. These results show that discrimination against both Black and male names is more intense in customer facing sectors, regardless of whether the job itself is customer facing. This finding may indicate that firms in different sectors tend to adopt different corporate cultures and human resources practices affecting all their jobs.

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 4 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.[12] It is worth noting, however, that many jobs in our sample did not specify an exact establishment address; consequently, 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 appears to be heightened 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 4, 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. Despite achieving statistical significance, these geographic correlations are all fairly modest in magnitude, which aligns with our earlier finding in Table 6 of a small but statistically significant between state variance component to racial discrimination.

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 4 shows that including firm fixed effects renders the establishment characteristics jointly insignificant ($p = 0.34$). Similar to our analysis of job titles, this finding 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. Consistent with our earlier finding in Table 6 of a negligible state component to gender gaps, Panel (b) of Figure 4 shows insignificant relationships between gender contact gaps and local demographics, measures of prejudice, and establishment charac-

[12]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.

teristics. 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 ($p \geq 0.35$).

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 5 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 5, 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.[13] 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 5, 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. As documented in Appendix Table C2 centralization is elevated among federal contractors ($p = 0.038$) but we cannot reject that it is unrelated to our other firm level predictors in a multivariate regression. Since this predictor is calculated using the outcome data 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,

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

1995) intended to avoid any mechanical relationship between job-level callback propensities and gaps. The negative coefficient estimate suggests that firms at which hiring responsibility is more centralized are less prone to bias, perhaps because rules replace the discretionary judgements of individual workers at firms with more sophisticated human resources practices. Overall, the firm-level variables in Figure 5 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 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 evaluating the sample average

$$\hat{g}_\Delta(x) = \frac{1}{108}\sum_{f=1}^{108}\frac{1}{s_f}\hat{g}_\mu(x/s_f).$$

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 and statistically insignificant relationships between standard errors and z-scores for both race and gender, suggesting that independence is a reasonable approximation. Appendix E explores three alternate approaches to modeling the joint distribution of s_f and z_f and shows that they yield results similar to those found when independence is imposed.

Panel (a) of Figure 6 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.[14] 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.[15] For comparison with the estimated densities, the background of Figure 6 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. Pointwise confidence intervals on the estimated densities are reported in Appendix Figure A10.

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

[14] 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.

[15] 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 A12.

discrimination is driven by a small group of highly discriminatory firms. Figure 7 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 contact gaps are even more concentrated, with firms in the top quintile responsible for 56% 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 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.[16]

10 Firm-Specific Estimates

The finding that discrimination is highly concentrated raises the question of whether it is possible to deduce the contact gaps of particular firms. Firm-specific estimates could, in principle, be shared with company executives, providing them with an assessment of their organization's biases, or with regulators to help them target audits or other enforcement efforts more effectively. Although the sample contact gaps $\hat{\Delta}_f$ provide unbiased estimates of the contact gap at each firm, those estimates are often quite noisy. 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 each specific firm.

10.1 Posterior mean 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

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

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 A11 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.

Appendix Figure A13 assesses the out of sample predictive power of these posterior means by shrinking z-scores constructing using only the first three waves of the experiment and comparing these shrunk values to contact gaps in the final two waves of the experiment. For race, we find a correlation between our EB predictions and the latent

contact gaps in the last two waves of 0.7, indicating very significant out of sample forecasting ability even when working with predictions that discard 40% of our microdata.

The posterior mean racial contact gaps vary systematically across industries. Figure 8 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 6, 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 9 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. These gender bias correlations align closely with the matched pair audit evidence reported by Neumark, Bank and Van Nort (1996) who find that women are discriminated against at upscale restaurants, which tend to pay high wages and to be male dominated, but are weakly preferred at lower price restaurants that tend to pay lower wages and to be female dominated.

One potential explanation for the divergent correlation patterns uncovered for sex and race in Figure 9 is that job seekers know that certain sectors (e.g. women's apparel) discriminate on the basis of gender, perhaps due to a mix of coworker and customer discrimination. This common knowledge allows workers to sort away from biased jobs, mitigating to some extent the burden of discrimination as in Becker's (1957) classic

model. Industry patterns of racial discrimination, by contrast, may be more difficult to discern, particularly if these patterns are driven by variation in opaque corporate recruiting protocols. When discriminatory patterns are not common knowledge, less pronounced sorting patterns will arise and a larger burden may fall 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 with degrees of freedom equal to the number of jobs at the firm minus one. Histograms of the resulting p-values for the null that firm specific contact gaps equal zero appear in Figure 10. 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 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. $FDR(p)$ 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

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

where $\widehat{FDR}(p)$ is an estimator of false discovery rates based on the empirical distribution of p-values.[18] If $\widehat{FDR}(p)$ were a consistent estimator of $FDR(p)$ 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 $FDR(p)$ 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

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

where $\lambda \in [0, 1)$ is a tuning parameter governing how deep to look in the right tail of empirical p-values. For any choice of λ, however, the probability limit of $\hat{\pi}_0(\lambda)$ will lie weakly above the true 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. To assess the degree of uncertainty in our estimate, we also report the upper limit of a non-parametric confidence interval for π_0 developed by Armstrong (2015).

[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 $\min_{t\geq\hat{p}_f} \widehat{FDR}(t)$, which ensures that q-values are non-decreasing for nested rejection thresholds.

10.3 Which firms discriminate?

Figure 10 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 10 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 $\hat{q}_f$.[19]

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.[20] 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 40% of firms may be discriminating against Black applicants, while under two-tailed tests the share discriminating may be as low as 30%.

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). The largest q-value in this set of firms is 0.047, so we should expect at most $23 \times 0.047 = 1.08$ false discoveries if these 23 firms are 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.

Though we expect at most 1 of the 23 firms with q-values below 0.05 to have racial contact gaps equal to zero, the actual number of false discoveries may differ from its expected value. To get a sense of how many false discoveries could potentially arise in an unfavorable scenario, Appendix Figure A14 plots the posterior distribution of false discoveries implied by the LFDRs of these 23 firms.[21] Reassuringly, the posterior probability mass function of false discoveries is tightly concentrated around its mean, with the

[19] 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.

[20] 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.

[21] The number of false discoveries exhibits a Poisson binomial posterior distribution with probabilities given by the LFDRs of the hypotheses under consideration. See Basu et al. (2021) for discussion.

posterior chances of three or more of these firms exhibiting contact gaps of zero being less than 2%.

The lower panels of Table 8 reveal that conclusions regarding the set of firms likely to be discriminating against Black names are remarkably robust to the method used to bound π_0. In fact, if we use randomization inference based p-values to estimate π_0, 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 8, 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 0.13, implying that at least 87% 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 A15 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 10 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 not close to one, a condition which seems to be violated 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

[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 $\hat{\pi}_0 = 1$. 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 $\phi_f < 1$.

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.

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 0.073. The mean gap across jobs is 0.020 with associated standard error of 0.002. Consequently, the lower bound prevalence is estimated to be $\frac{(0.020)^2-(0.002)^2}{(0.020)^2-(0.002)^2+(0.073)^2} \approx 0.07$, 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 12 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 $\hat{q}_f < 0.05$ 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 employer 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]

[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 as 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 $p \to \infty$, the auditor grows increasingly risk averse and U approaches U^e.

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 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 $\bar{\Delta}_f(G) = \mathbb{E}_G[\Delta_f|\mathcal{E}]$ 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 11 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. Since $\hat{G}_\Delta$ exhibits a mean contact gap of 2.1 percentage points (see Figure 6), 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 approximately 37% of lost contacts. Investigating the same share of firms based on posterior mean rankings would be expected to yield only an additional 4% of 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 the contacts lost to racial discrimination while maintaining a tight limit on the expected number of 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 concentration of discrimination among particular employers may amplify group disparities if discriminatory firms tend to offer higher wages. While we found no relationship at the industry level between racial wage gaps and racial contact gaps, industry contact gaps favoring men were found to be predictive of larger gender wage gaps. An interesting topic for future research is to assess the extent to which the firm-level contact gaps uncovered in this experiment correlate with group disparities in firm wage fixed effects such as those studied by Card, Cardoso and Kline (2016) or Gerard et al. (2021).

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. 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 EEOC maintains an internal target for the share of its litigation docket comprised of systemic discrimination cases. The appropriate level of this target has been a topic of recurring debate in Congress, where oversight committees have questioned the EEOC's

choice to prioritize systemic investigations of firms over individual level claims of discrimination (Kim, 2015). Our finding that discrimination is highly concentrated in particular companies lends some credence to the notion that appropriately targeted systemic investigations have the potential to remedy, and perhaps also prevent, discrimination affecting a wide swath of the labor force. Whether the targeting of systemic investigations by the EEOC, or of contractor audits by the OFCCP, can be improved with the adoption of statistical methods of the sort developed here is an open and interesting question.

Enforcement actions are inevitably costly and contentious. It is natural to wonder whether bias at the most discriminatory firms can be preemptively reduced or eliminated by modifying 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 (negative) predictors of both racial and gender contact gaps found in our correspondence experiment is callback centralization, which is notably elevated among federal contractors. This finding leads us to suspect that human resources practices play an important role in translating the biased judgements of individuals into biased behavior by organizations. While centralizing interview decisions might serve to reduce discrimination, such changes may also simply postpone discrimination to a later stage of the hiring 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).

[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."

References

Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. "Teachers and student achievement in the Chicago public high schools." *Journal of labor Economics*, 25(1): 95–135.

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.

Basu, Pallavi, Luella Fu, Alessio Saretto, and Wenguang Sun. 2021. "Empirical bayes control of the false discovery exceedance." *arXiv preprint arXiv:2111.03885*.

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.

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

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." *University of Chicago Legal Forum*, 139.

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.

Gerard, François, Lorenzo Lagos, Edson Severnini, and David Card. 2021. "Assortative matching or exclusionary hiring? the impact of employment and pay policies on racial wage differences in brazil." *American Economic Review*, 111(10): 3418–57.

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*.

Hangartner, Dominik, Daniel Kopp, and Michael Siegenthaler. 2021. "Monitoring hiring discrimination through online recruitment platforms." *Nature*, 589(7843): 572–576.

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.

Hurst, Erik, Yona Rubinstein, and Kazuatsu Shimizu. 2021. "Task-based discrimination." 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.

Kim, Pauline T. 2015. "Addressing systemic discrimination: Public enforcement and the role of the EEOC." *BUL Rev.*, 95: 1133.

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.

Koenker, Roger, and Ivan Mizera. 2014. "Convex optimization, shape constraints, compound decisions, and empirical Bayes rules." *Journal of the American Statistical Association*, 109(506): 674–685.

Koenker, Roger, and Jiaying Gu. 2017. "REBayes: an R package for empirical Bayes mixture methods." *Journal of Statistical Software*, 82: 1–26.

Krueger, Alan B, and Lawrence H Summers. 1988. "Efficiency wages and the inter-industry wage structure." *Econometrica: Journal of the Econometric Society*, 259–293.

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

Maxwell, Nan, Aravind Moorthy, Caroline Massad Francis, Dylan Ellis, et al. 2013. "Using Administrative Data to Address Federal Contractor Violations of Equal Employment Opportunity Laws." Mathematica Policy Research.

Miller, Conrad. 2017. "The persistent effect of temporary affirmative action." *American Economic Journal: Applied Economics*, 9(3): 152–90.

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.

Neumark, David, Roy J Bank, and Kyle D Van Nort. 1996. "Sex discrimination in restaurant hiring: An audit study." *The Quarterly journal of economics*, 111(3): 915–941.

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_2016OFC00006_(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.

Pincus, Fred L. 1996. "Discrimination comes in many forms: Individual, institutional, and structural." *American Behavioral Scientist*, 40(2): 186–194.

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.

Reskin, Barbara. 2012. "The race discrimination system." *Annual Review of Sociology*, 38: 17–35.

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.

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."

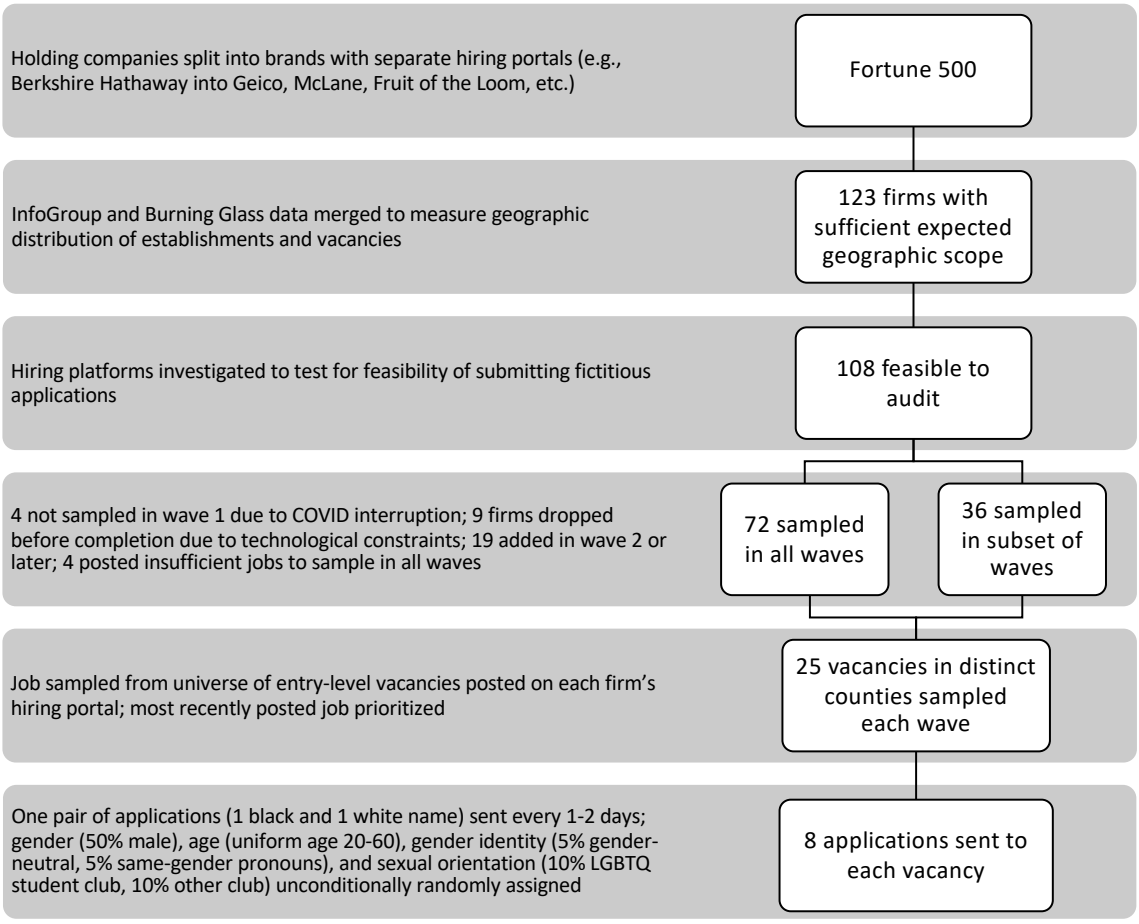
U.S. EEOC. 2019. "EEOC Sues Schuster for Sex Discrimination." *Press Release*, https://www.eeoc.gov/newsroom/eeoc-sues-schuster-sex-discrimination.

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.

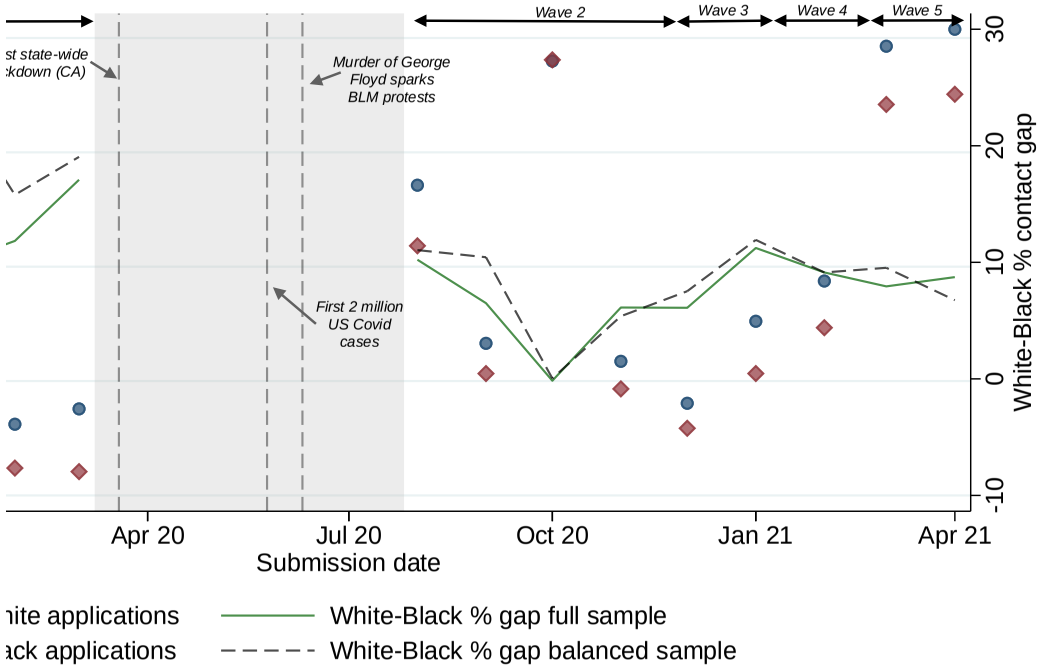
Figures

Figure 1: Overview of sampling strategy and experimental design



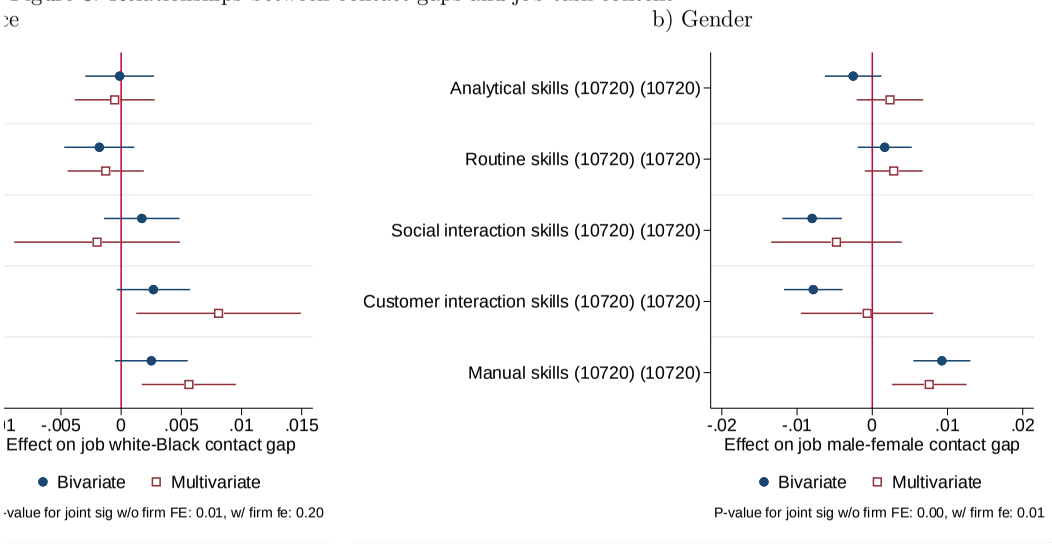
Notes: This figure explains the sampling strategy and design for the experiment. Gender identity and sexual orientation attributes were assigned starting in wave 2 after the U.S. Supreme Court ruling in Bostock v. Clayton County, Georgia.

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



t rates for black and white applications by month and year of submission. The gray region corresponds to the ased due to Covid-19 related shut-downs. The green and dotted black lines plot the white-Black contact rate ack contact rate for the full and balanced samples. An F-test fails to reject that the white/Black percentage e same in all waves of the experiment ($F = 0.85$, $p = 0.50$).

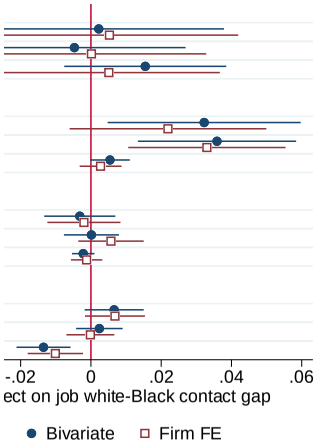
Figure 3: Relationships between contact gaps and job task content



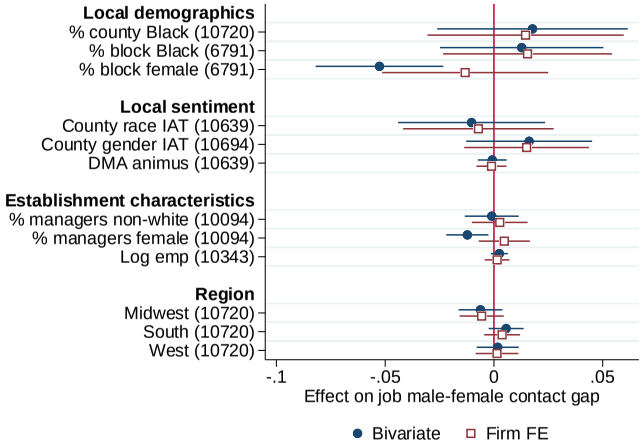
hip between O*Net measures of job-level task content and contact gaps for race and gender. Each relationship with job-level contact gaps as the outcome. All jobs with defined contact gaps for each attribute are included. on is in parentheses. Task measures are normalized to have standard deviation one in sample. "Bivariate" ons of contact gaps on the covariate alone. "Multivariate" points plot effects when all covariates are included confidence intervals based on robust standard errors. Appendix C provides a complete description of task

e 4: Relationships between contact gaps and establishment characteristics

ce

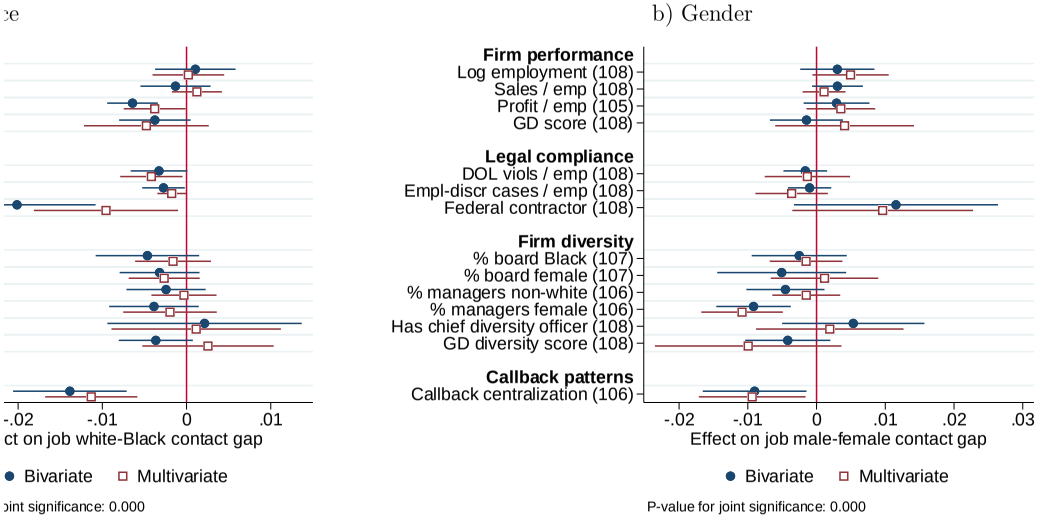


b) Gender



hip between establishment-level covariates and contact gaps for race and gender. Each relationship is estimated l contact gaps as the outcome. All jobs with defined contact gaps for each attribute and matched to the listed of jobs in each regression is in parentheses. Covariates are standardized to be mean zero, standard deviation 1 efficients from regressions of contact gaps on the covariate alone. "Firm FE" points include firm fixed effects. als based on robust standard errors. The omitted region category is the Northeast. Appendix C provides a initions and sources.

Figure 5: Relationships between contact gaps and firm characteristics

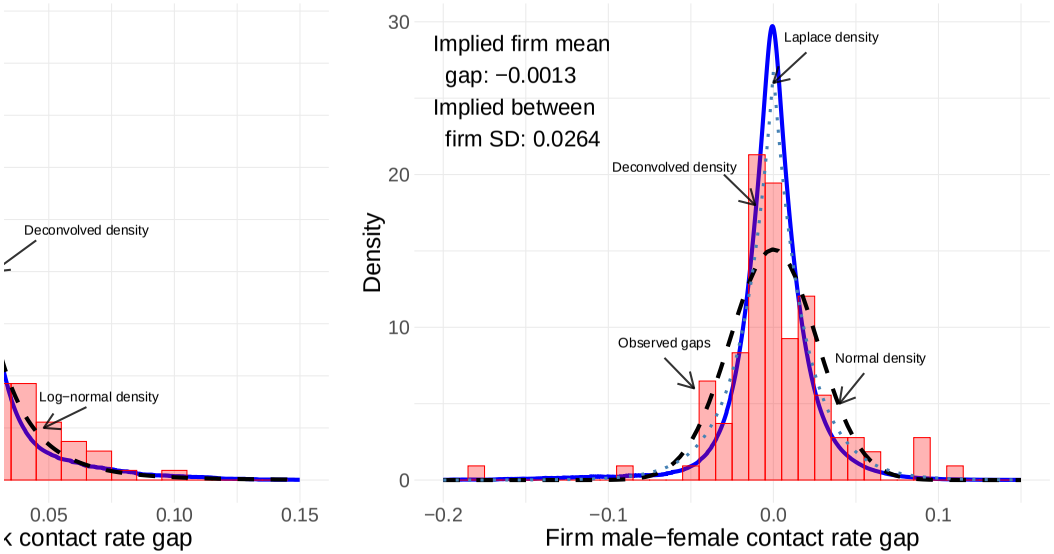


ship between firm-level covariates and contact gaps for race and gender. Each relationship is estimated by a ct gaps as the outcome, except for callback centralization which is estimated via split sample IV to account for outcome. The number of firms in each regression is in parentheses. Covariates are standardized to be mean , except for the binary indicators for federal contractor status and having a chief diversity officer. "Bivariate" ons of contact gaps on the covariate alone. "Multivariate" points plot effects when all covariates are included lization measured in half of the randomly split sample instrumented using its value in the other half and all ous regressors. Bars indicate 95% confidence intervals based on standard errors clustered at the firm level. cription of covariate definitions and sources.

ure 6: Deconvolution estimates of firm-level discrimination distributions

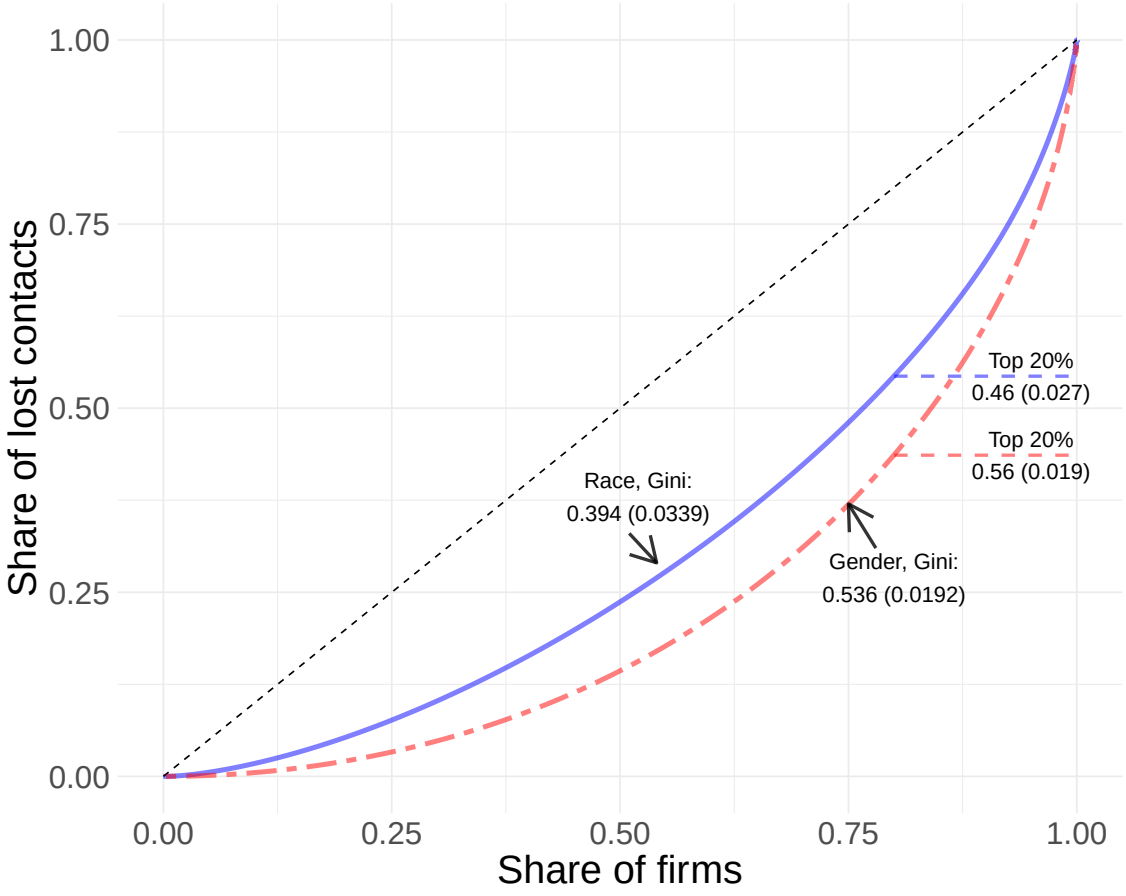
ce

b) Gender



metric estimates of the distribution of firm-specific contact gaps. Panel (a) presents estimates for white-Black) presents estimates for male-female differences. Red histograms show the distribution of estimated firm contact mates of population contact gap distributions. The population distributions are estimated by applying the nd Efron, 2020) to firm-specific z-score estimates, then numerically integrating over the empirical distribution stribution of contact gaps. The penalization parameter in the deconvolution step is calibrated so that the orresponding bias-corrected variance estimate from Table 4. In panel (a), the density of population z-scores Parametric comparison distributions shown with dashed and dotted lines are calibrated to have means and nvolved distributions.

Figure 7: Discrimination Lorenz curves



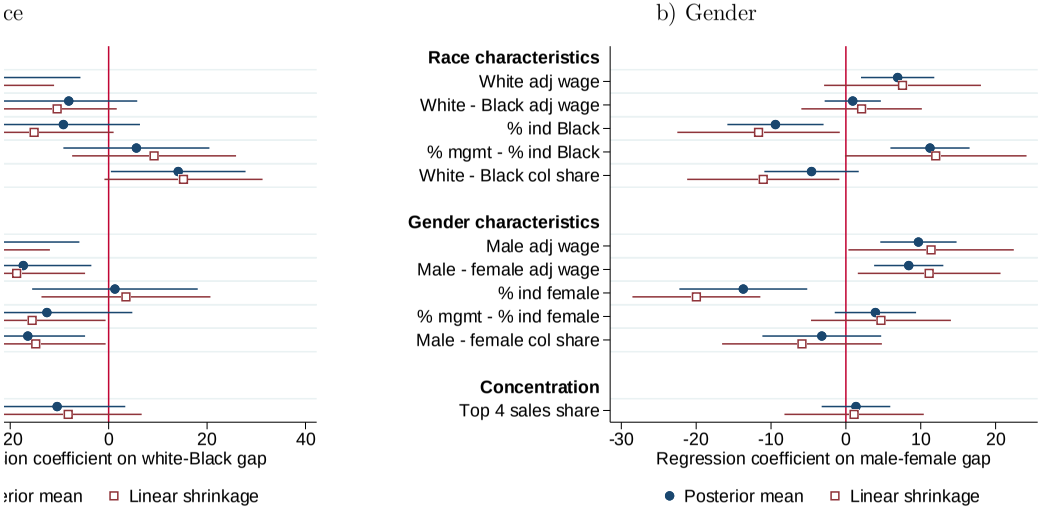
Notes: This figure displays Lorenz curves implied by the non-parametric deconvolution estimates of race and gender contact gap distributions in Figure 6. The solid blue curve is the Lorenz curve for the white/Black contact gap, and the dashed red curve is the Lorenz curve for the absolute value of the 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. Standard errors for Gini coefficients and top 20% shares are reported in parentheses. Standard errors are computed using 1,000 iterations of a parametric bootstrap redrawing a bootstrap observation $\hat{\Delta}_{fb}$ for each firm from a $\mathcal{N}(\hat{\Delta}_f, s_f^2)$ distribution.

Figure 8: Posterior means by industry



y-level averages of posterior mean contact gaps. The navy bars show average posteriors using deconvolved gaps as priors. The light red bars show averages of estimates shrunk linearly towards the grand mean with ratio $\hat{\theta}/(s_f^2 + \hat{\theta})$.

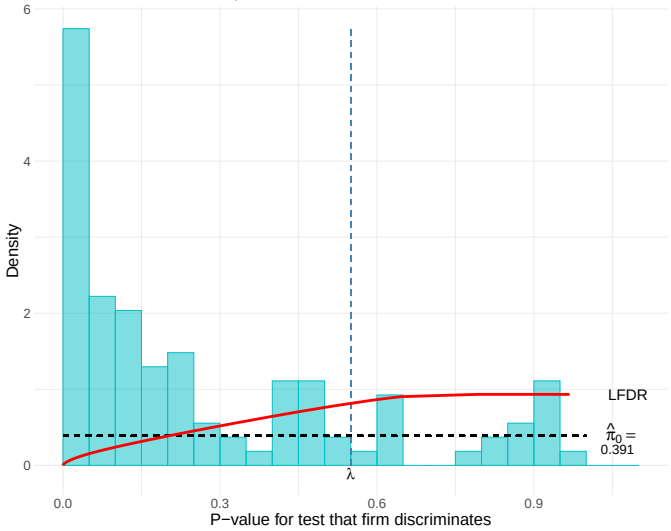
Figure 9: Industry correlates of contact gaps



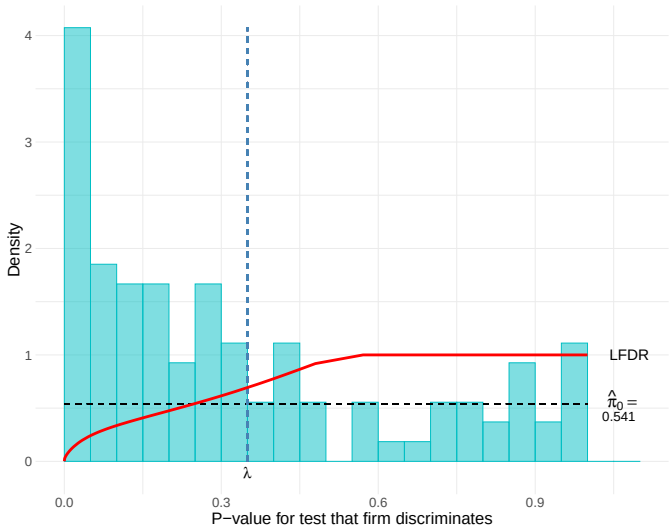
ns of industry characteristics on posterior mean contact gaps for race and gender. Points labeled "posterior r gaps formed using the distributions in Figure 6 as priors. Points labeled "linear shrinkage" show coefficients grand mean with weights given by the signal to noise ratio $\hat{\theta}/(s_f^2 + \hat{\theta})$. Outcomes are normalized to be mean . Appendix C provides a complete description of covariate definitions and sources.

Figure 10: *P*-value distributions and local false discovery rates

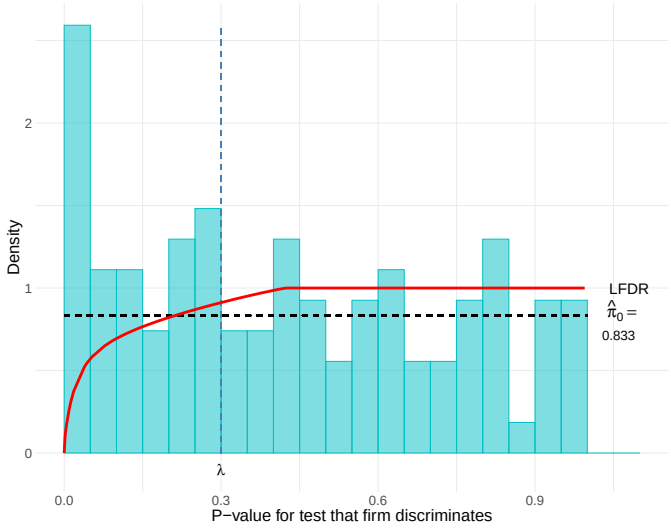
a) Race, one-sided



b) Race, two-sided

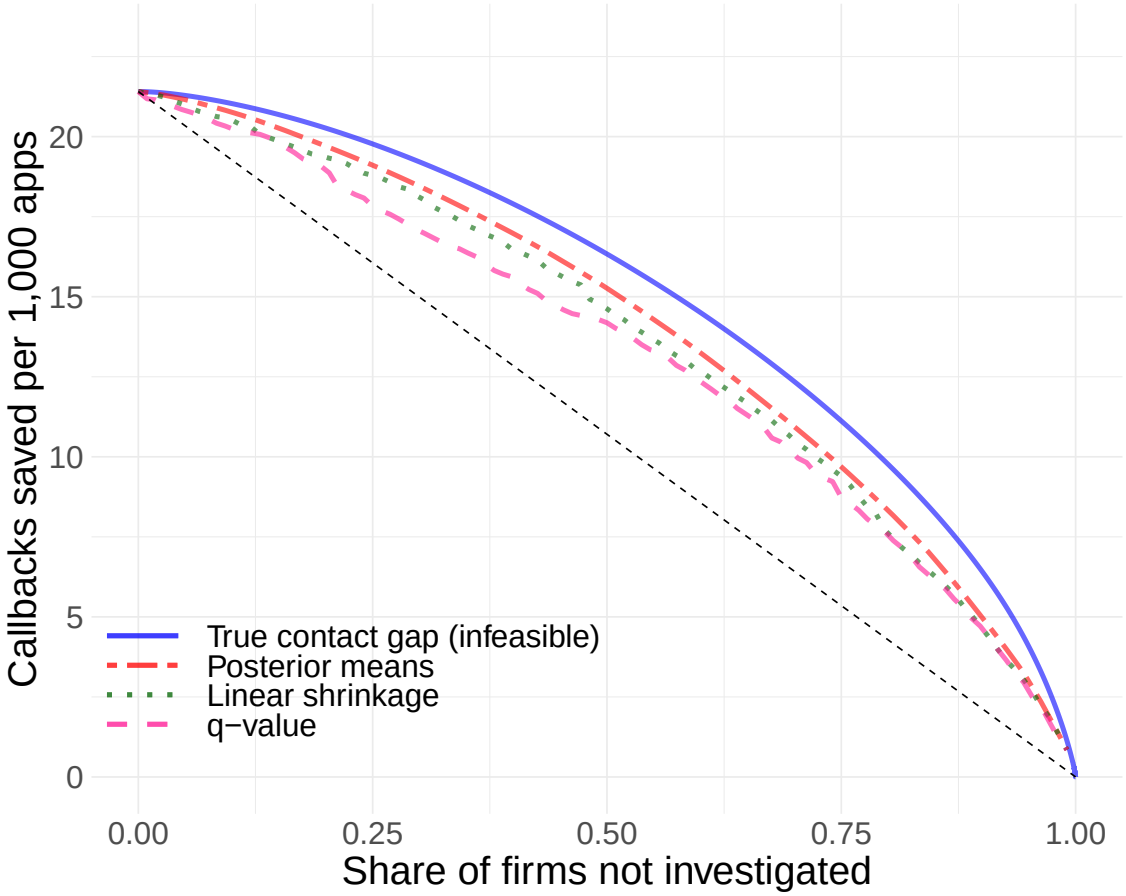


c) Gender, two-sided



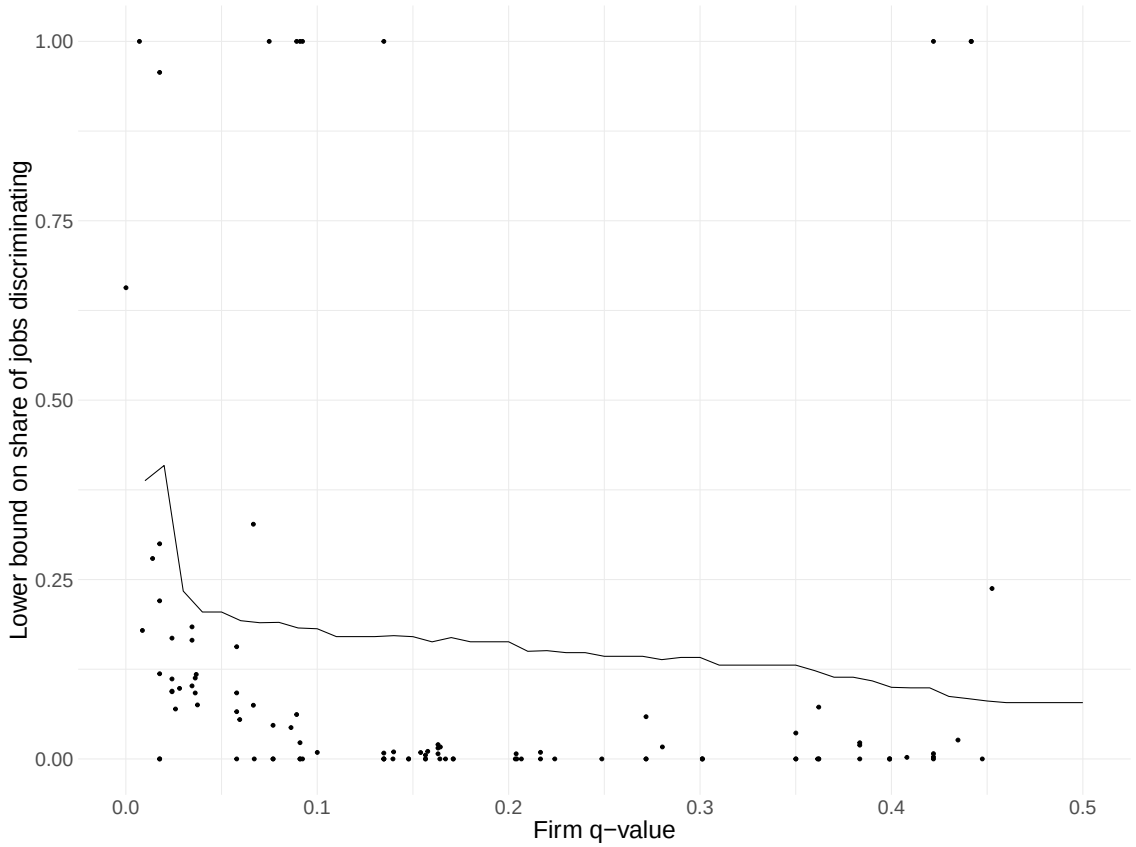
Notes: This figure plots distributions of *p*-values from firm-specific tests of the null hypothesis of no discrimination. Panel (a) shows results for one-sided tests of no discrimination against Black applicants, and panel (b) displays results for two-sided tests of equal contact rates for Black and 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 11: 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 12: 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			3/4/14/15/72			

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) LPM	(2) Logit	(3) LPM	(4) 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	-0.0000389	-0.000671
	(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	83643	83643	65368	65368

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

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

	OLS			Logit		
	(1) White	(2) Black	(3) Difference	(4) White	(5) Black	(6) 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)
Associate 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	41837	41806	83643	41837	41806	83643
χ^2 stat for joint significance			14.71			14.54
p-value			0.0650			0.0687

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

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

			Contact gap SD		
	(1) χ^2 test of heterogeneity	(2) p-value for no discrim against:	(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	(0.0031)	(0.0032)	(0.0031)
Gender	205.2	M: 0.00	0.0267	0.0287	0.0269
	[0.000]	F: 0.05	(0.0038)	(0.0035)	(0.0038)
Over 40	144.6	Y: 0.22	0.0103	0.0044	0.0086
	[0.011]	O: 0.02	(0.0069)	(0.0158)	(0.0082)

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
	(0.0036)	(0.0018)	(0.0276)	(0.0152)	(0.0238)	(0.0131)
Std. dev.	0.1607	0.0186	0.9755	0.1155	0.7047	0.0837
	(0.0035)	(0.0035)	(0.0385)	(0.0360)	(0.0382)	(0.0341)
Corr. w/own slope	-0.4010	1.000	0.0519	1.000	0.0685	1
	(0.1098)	-	(0.2074)	-	(0.3092)	-
Corr. w/LPM slope	-0.4010	1.000	-0.4274	0.8944	-0.5045	0.8075
	(0.1098)	-	(0.1068)	(0.2095)	(0.1149)	(0.3074)
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	-	-
	(0.0034)		
	[0.038]	[0.668]	[0.583]
Industry	0.0141	0.0190	0.0048
	(0.0021)	(0.0029)	(0.0053)
	[0.000]	[0.000]	[0.112]
Job title	0.0136	0.0111	0.0034
SOC3 code	(0.0025)	(0.0043)	(0.0105)
	[0.000]	[0.007]	[0.527]
Hiring platform	0.0059	0.0024	0.0024
intermediary	(0.0025)	(0.0088)	(0.0071)
	[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
	0.0176	0.0150	0.0253	0.0255	0.0096	0.0088
ects	0.0003	-	-	0.0080	0.0004	-
	0.0000	0.0001	0.0000	0.0002	0.0000	0.0002
	11026	11026	10720	10720	10652	10652
	108	108	108	108	108	108
	51	47	51	47	51	47
1 firm	51	43	51	43	51	43
	0.0196	0.0196	0.0023	0.0023	0.0037	0.0037
	0.000	0.0008	0.000	0.000	0.071	.040
e effects	0.186	0.327	0.482	0.237	0.86	0.459

cted variance component estimates from two-way fixed effect models estimated using the leave-out procedure of olumns labeled "Job title" include fixed effects for the first three digits of each job's O*Net SOC code. Columns for the job's state. All variance and covariance estimates are job-weighted. Only jobs in the leave-job-out stimate. Dashes indicate negative variance estimates and hence undefined estimated standard deviations. "N mber of states or job titles in the connected set observed at 2 or more firms. The final two rows report p-values at all firm or job title / state fixed effects equal zero, computed using the heteroscedasticity robust procedure

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			
$\hat{\pi}_0$	0.370	0.455	0.808	0.802
# q-values <= 0.05	35	24	8	1
# q-values <= 0.1	55	36	10	1
λ	0.550	0.450	0.450	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			
$\hat{\pi}_0$	0.602	0.696	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).

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

	Federal contractor?	Contact gap	Std. err.	p-value	q-value	Posterior mean	Posterior 5th pctile	Posterior 95th pctile
rvices	Yes	0.0952	0.0197	0.0000	0.0001	0.0835	0.0450	0.1035
rvices	No	0.0507	0.0143	0.0003	0.0061	0.0354	0.0135	0.0673
rvices	No	0.0738	0.0220	0.0005	0.0073	0.0489	0.0192	0.0981
rvices	No	0.0787	0.0249	0.0010	0.0103	0.0498	0.0202	0.1031
es	No	0.0733	0.0250	0.0022	0.0158	0.0448	0.0187	0.0929
.	No	0.0469	0.0159	0.0020	0.0158	0.0286	0.0119	0.0595
.	Yes	0.0605	0.0219	0.0033	0.0176	0.0365	0.0154	0.0743
adise	Yes	0.0520	0.0187	0.0031	0.0176	0.0314	0.0132	0.0641
rvices	No	0.0613	0.0240	0.0060	0.0194	0.0370	0.0158	0.0725
.	No	0.0560	0.0214	0.0050	0.0194	0.0337	0.0143	0.0669
ng	No	0.0560	0.0222	0.0064	0.0194	0.0339	0.0144	0.0660
rvices	No	0.0540	0.0215	0.0068	0.0194	0.0327	0.0139	0.0634
	Yes	0.0511	0.0204	0.0069	0.0194	0.0310	0.0132	0.0599
adise	No	0.0427	0.0170	0.0068	0.0194	0.0259	0.0110	0.0502
res	Yes	0.0400	0.0159	0.0066	0.0194	0.0242	0.0103	0.0470
rable	No	0.0386	0.0158	0.0080	0.0199	0.0235	0.0100	0.0450
turing	Yes	0.0350	0.0142	0.0078	0.0199	0.0213	0.0090	0.0409
ials	Yes	0.0373	0.0157	0.0093	0.0218	0.0229	0.0097	0.0433
es	Yes	0.0544	0.0240	0.0132	0.0292	0.0339	0.0143	0.0627
res	No	0.0400	0.0183	0.0152	0.0322	0.0252	0.0106	0.0460
ng	No	0.0340	0.0159	0.0172	0.0346	0.0217	0.0090	0.0392
adise	No	0.0423	0.0210	0.0229	0.0439	0.0277	0.0114	0.0494
estate	No	0.0278	0.0140	0.0257	0.0472	0.0183	0.0075	0.0325

f white-Black contact gaps for the 23 individual firms with q-values less than 0.05. P-values and q-values come thesis that the firm does not discriminate against Black applicants. To ensure that q-values are non-decreasing low 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 ans and percentiles are empirical Bayes posteriors constructed using the estimated distribution in Figure 6 as

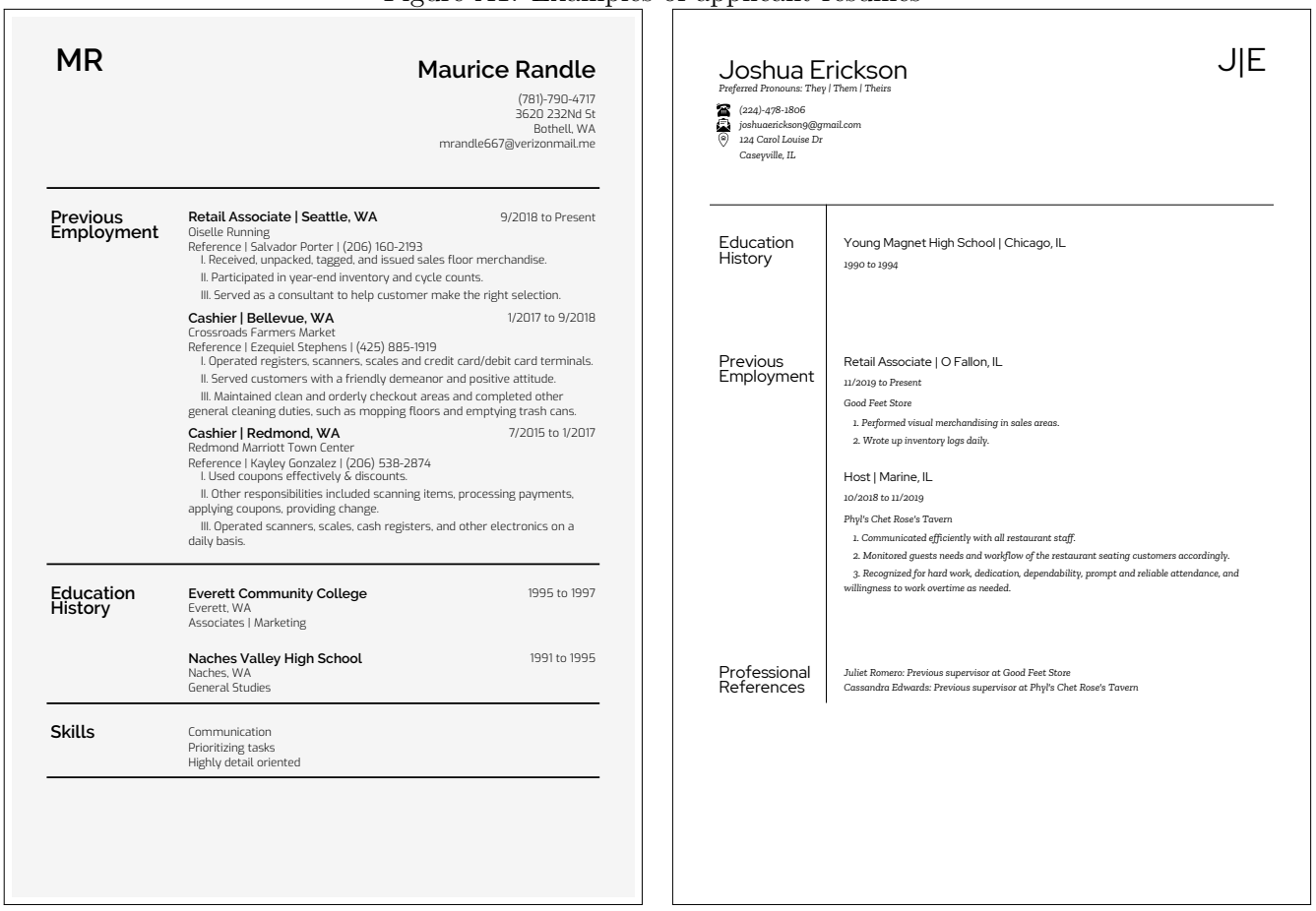
Table 10: Discrimination estimates and detection by industry

SIC	Industry	N firms	Race			Gender		
			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	0.900	-0.004	0	0.993
23	Apparel manufacturing	2	0.021	1	0.170	0.007	0	0.702
24-35	Other manufacturing	4	0.018	0	0.361	0.012	0	0.669
42-47	Freight / transport	4	0.011	0	0.822	0.001	0	0.941
48	Communications	2	0.017	0	0.340	0.013	0	0.972
49	Electric / gas	3	0.015	0	0.339	0.002	0	0.980
50	Wholesale durable	2	0.017	0	0.293	0.034	0	0.555
51	Wholesale nondurable	11	0.018	1	0.456	0.005	0	0.865
52	Building materials	3	0.014	1	0.544	0.012	0	0.849
53	General merchandise	12	0.023	3	0.276	-0.001	0	0.867
54	Food stores	5	0.025	1	0.356	0.009	0	0.821
55	Auto dealers / services	8	0.040	6	0.127	0.005	0	0.882
56	Apparel stores	4	0.025	1	0.253	-0.061	1	0.416
57	Furnishing stores	4	0.022	2	0.304	-0.006	0	0.787
58	Eating/drinking	5	0.027	2	0.303	0.003	0	0.926
59	Other retail	7	0.022	3	0.314	-0.002	0	0.971
60-61	Banks / credit	5	0.010	0	0.651	0.002	0	0.778
62	Securities brokers	2	0.010	0	0.410	-0.011	0	0.654
63-65	Insurance / real estate	8	0.013	1	0.463	-0.003	0	0.915
70	Accommodation	2	0.015	0	0.527	0.001	0	1.000
73	Business services	3	0.012	0	0.539	0.000	0	0.942
75-76	Auto / repair services	3	0.013	0	0.474	0.015	0	0.624
80	Health services	5	0.016	1	0.726	-0.009	0	0.909
87	Engineering services	2	0.009	0	0.348	-0.001	0	0.965

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 10. The distribution of gender LFDRs is depicted in Panel (c) of Figure 10.

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
joshuaericksn9@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.*

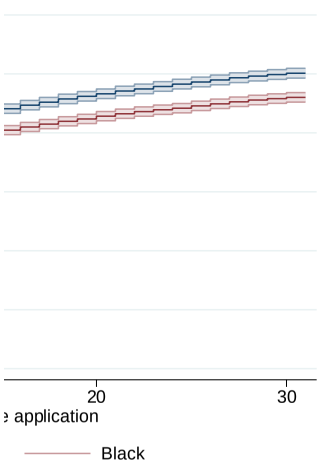
Professional References

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

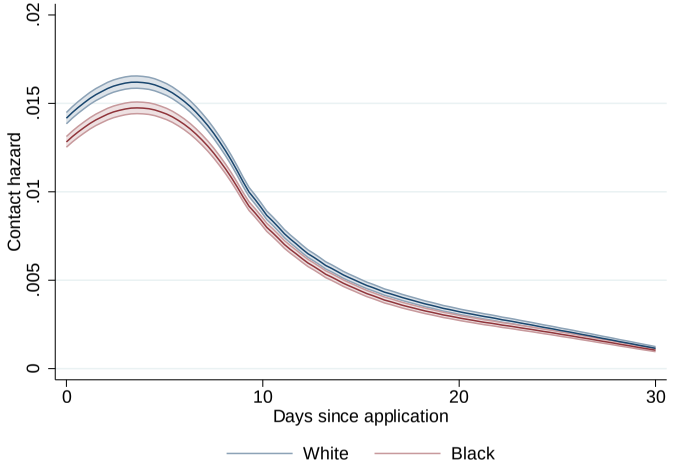
Notes: This figure presents two examples of randomly generated resumes used in the experiment. Resumes are formatted using a combination of pre-set options specifying length, fonts, text sizes, section header names, and layouts, with controls to ensure resumes that overflow one page are not generated.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

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

robability

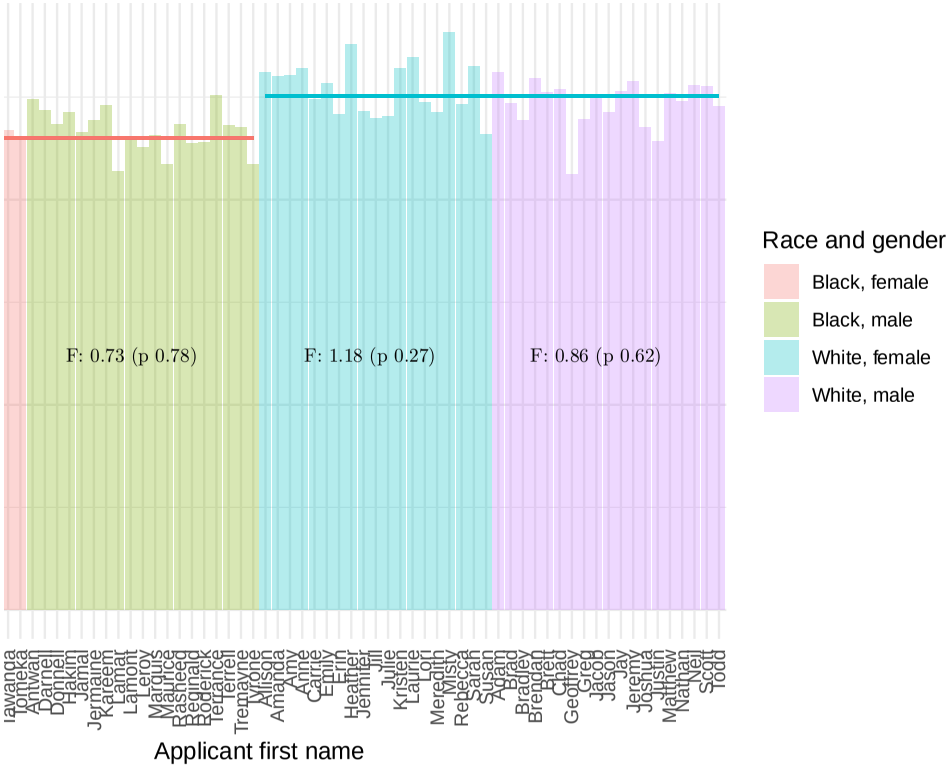


b) Smoothed contact hazard



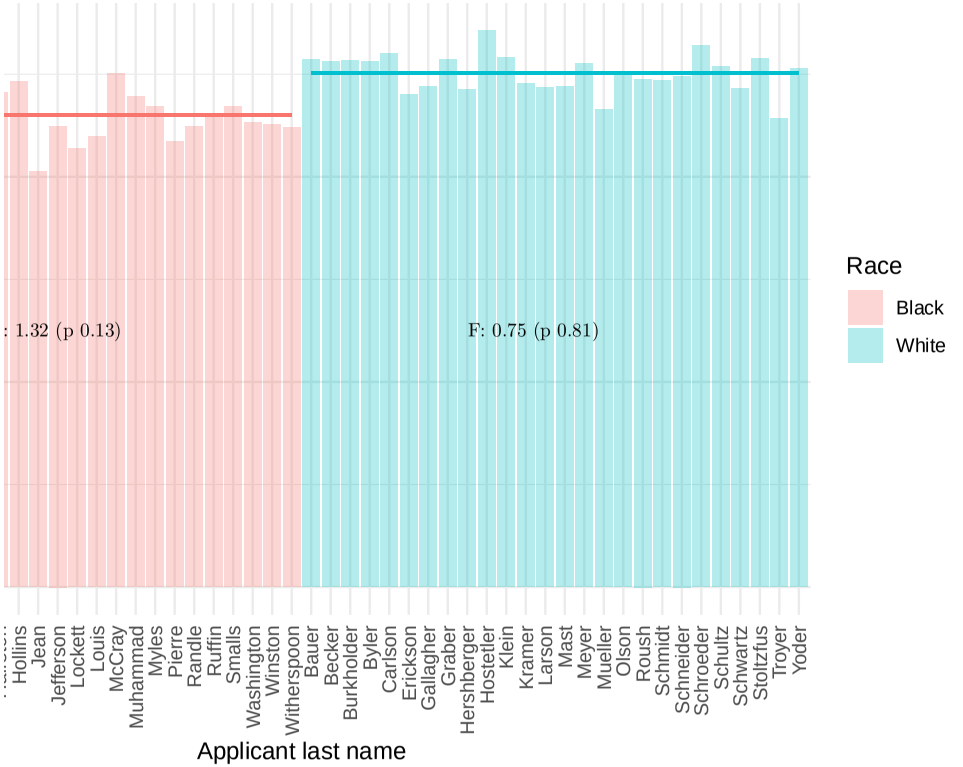
babilities and hazards as functions of days since application. Contact probabilities correspond to Kaplan-Meier are Kaplan-Meier hazard estimate smoothed using the Epanechnikov kernel. Shaded areas represent pointwise

Figure A3: Callbacks by applicant first name



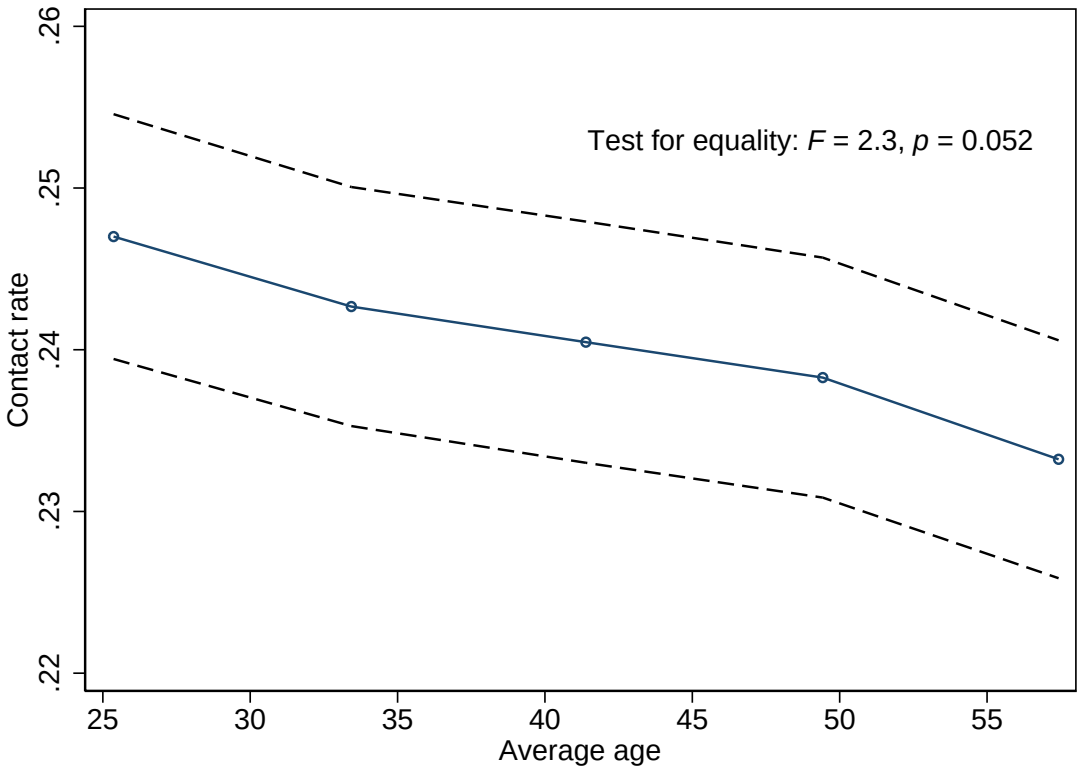
act rates by applicant first name, organized by race and gender group. The horizontal bars show race group
alues come from joint tests of the hypothesis that contact rates are equal across names separately by race and

Figure A4: Callbacks by applicant last name



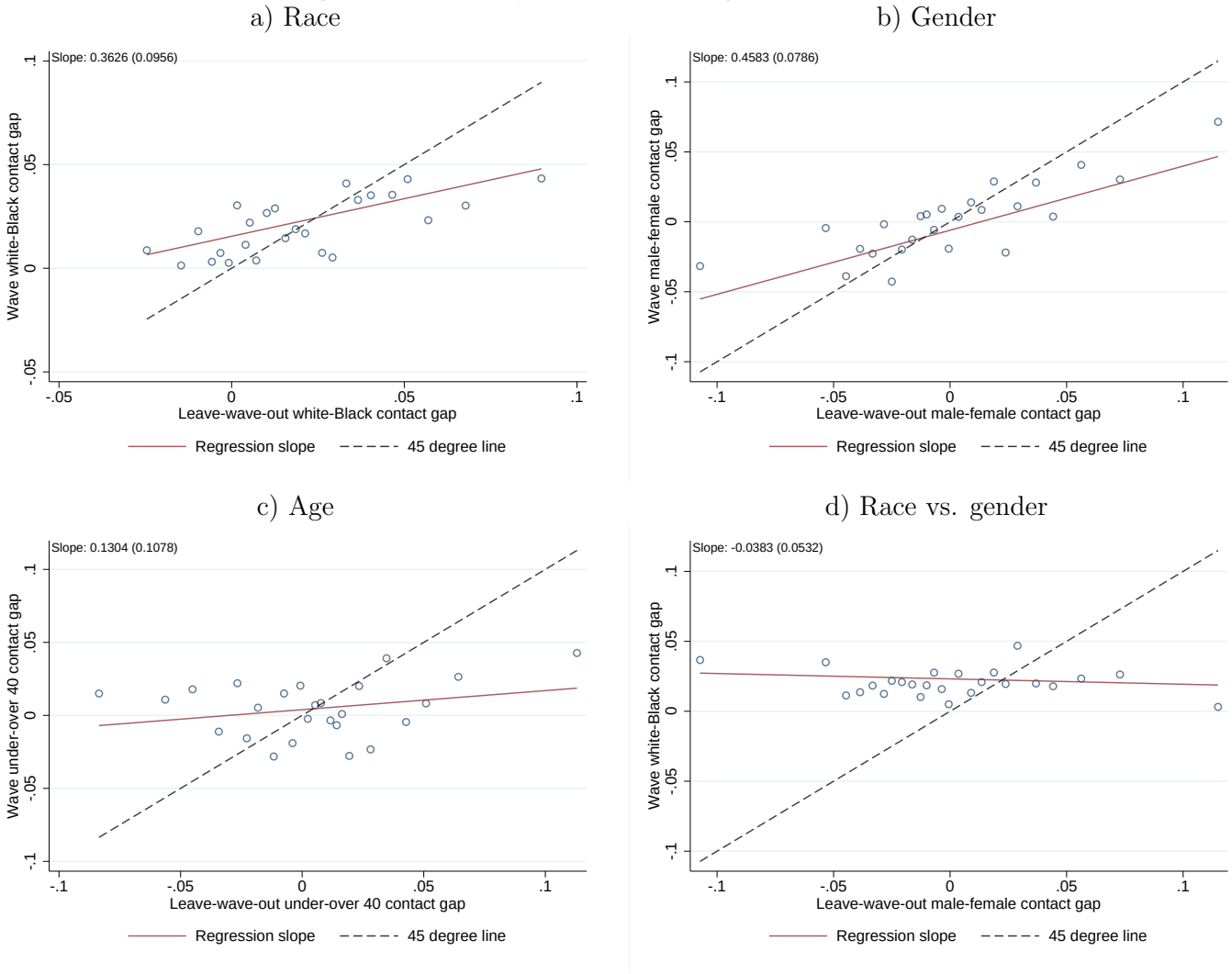
ıct rates by applicant last name, organized by race. The horizontal bars show race group mean contact rates.
: tests of the hypothesis that contact rates are equal across names separately by race.

Figure A5: Contact rates by age category



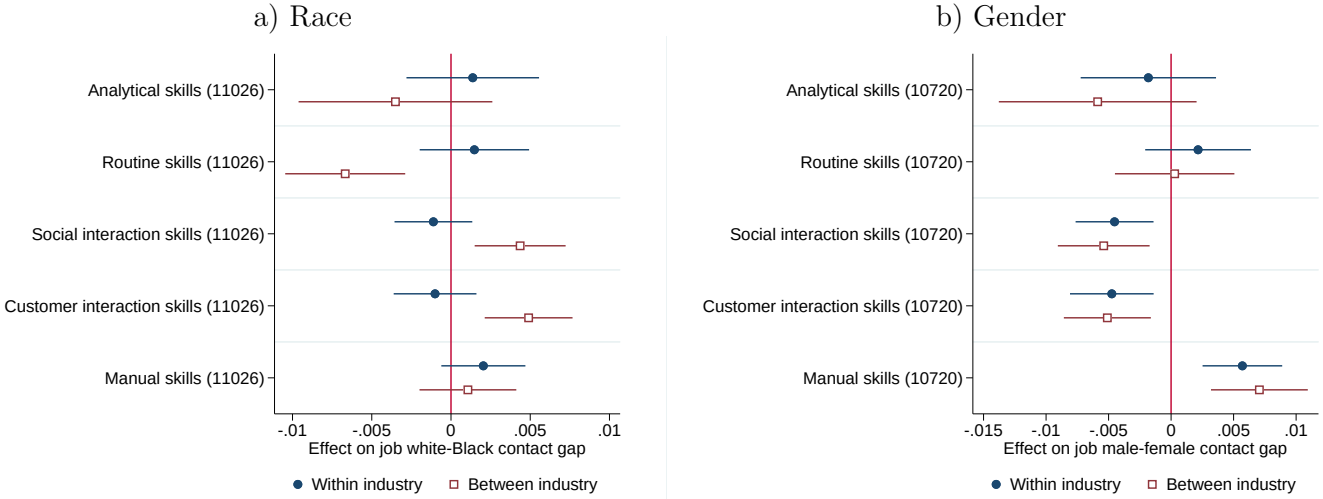
Notes: This figure plots average 30-day contact rates by quintile of applicant age at the time of application. Estimates come from regressions of a contact indicator on indicators for age quintile, controlling for wave indicators. The horizontal axis plots average age in each quintile. The vertical axis 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: Stability of firm contact gaps across waves



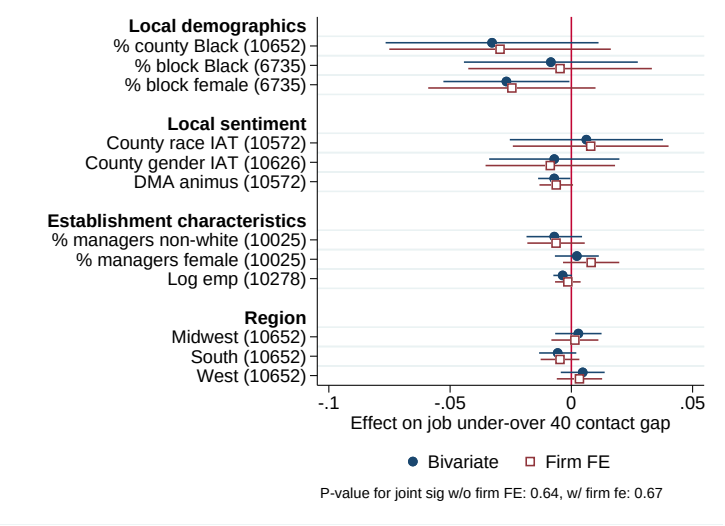
Notes: This figure presents binned scatter plots of firm-specific wave-average contact gaps vs. leave-wave-out firm-specific average contact gaps. Panel (a) reports results for the white/Black difference in contact rates. Panel (b) shows results for the male/female difference in contact rates. Panel (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 vingtiles 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 A7: Within and between industry relationship between contact gaps and task content



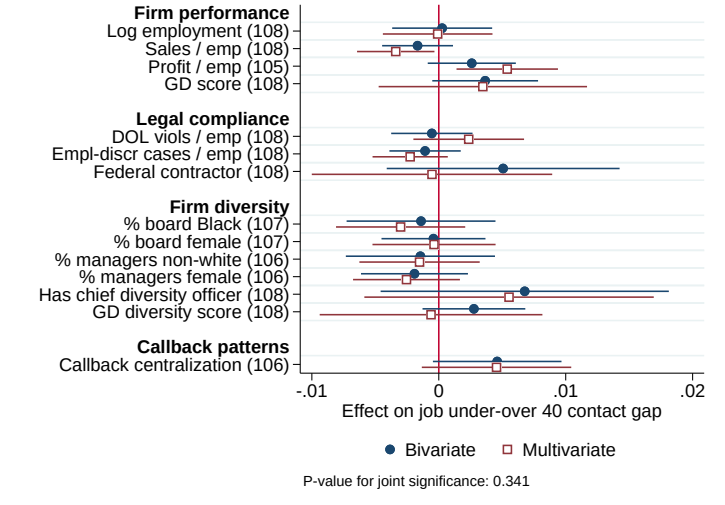
Notes: This figure plots the relationship between O*Net measures of job-level task content and contact gaps for race and gender within and between industry. The within relationship is estimated with a linear regression with job-level contact gaps as the outcome and two-digit industry fixed effects. The between relationship is estimated by instrumenting job task content with industry dummies. 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. Bars indicate 95% confidence intervals based on robust standard errors. Appendix C provides a complete description of task definitions and sources.

Figure A8: Relationships between age contact gaps and establishment characteristics



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.

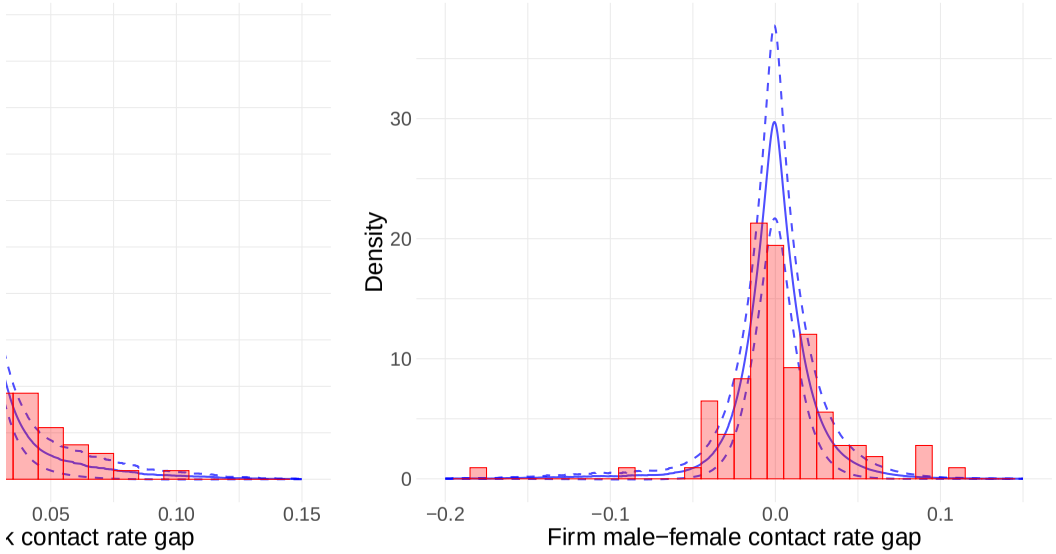
Figure A9: Relationships between age contact gaps and firm characteristics



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.

e A10: Confidence intervals on deconvolutions of firm-level discrimination

ce b) Gender

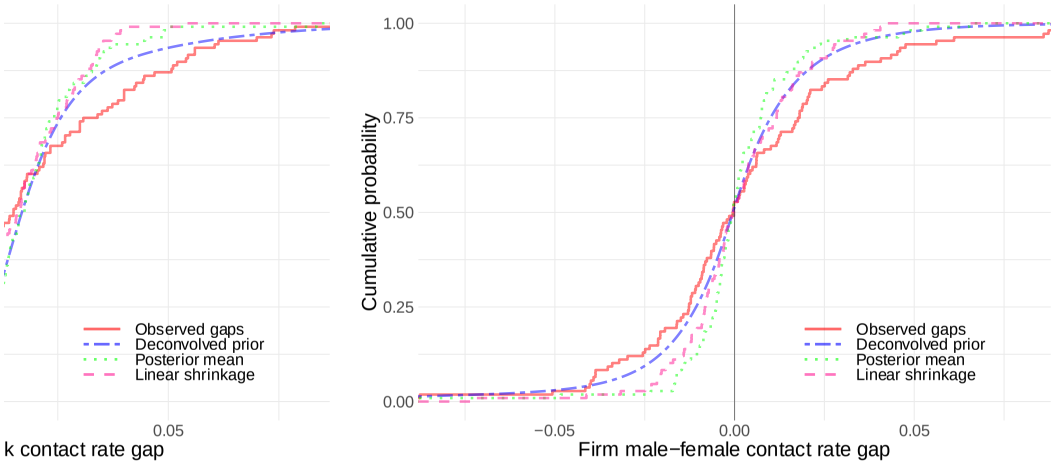


metric estimates of the distribution of firm-specific contact gaps and point-wise 95% confidence intervals. Panel k contact rate differences, and panel (b) presents estimates for male-female differences. Red histograms show ıtact gaps. Blue lines shows estimates of population contact gap distributions. The population distributions are eR package (Narasimhan and Efron, 2020) to firm-specific z-score estimates, then numerically integrating over d errors to recover the distribution of contact gaps. Since the estimated density of Δ is a linear combination rd errors can be computed using the delta method and the variance-covariance of $\hat{g}_\mu$ produced by Narasimhan parameter in the deconvolution step is calibrated so that the resulting distribution matches the corresponding ı Table 4. In panel (a), the density of population z-scores is constrained to be weakly positive.

ure A11: Distribution of observed, deconvolved, and posterior estimates

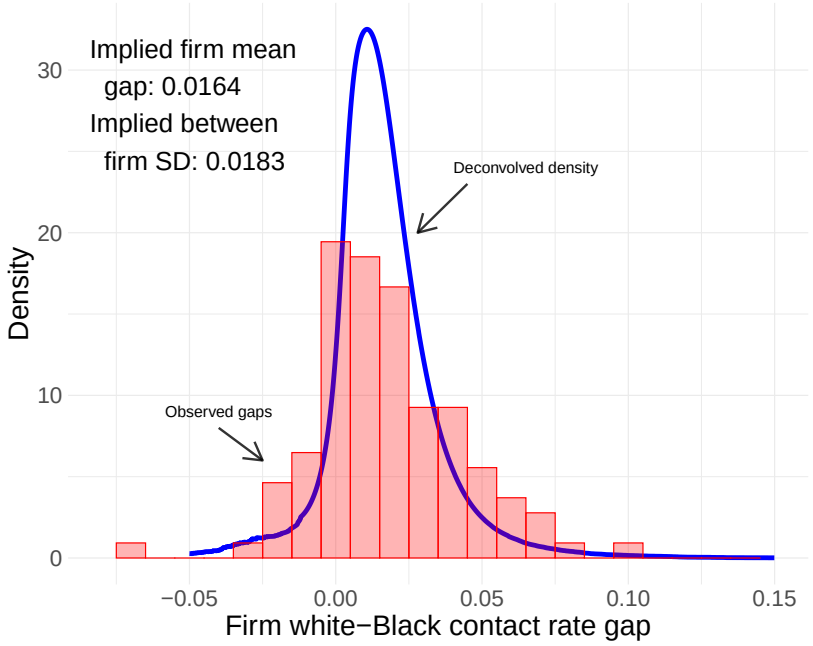
ce

b) Gender



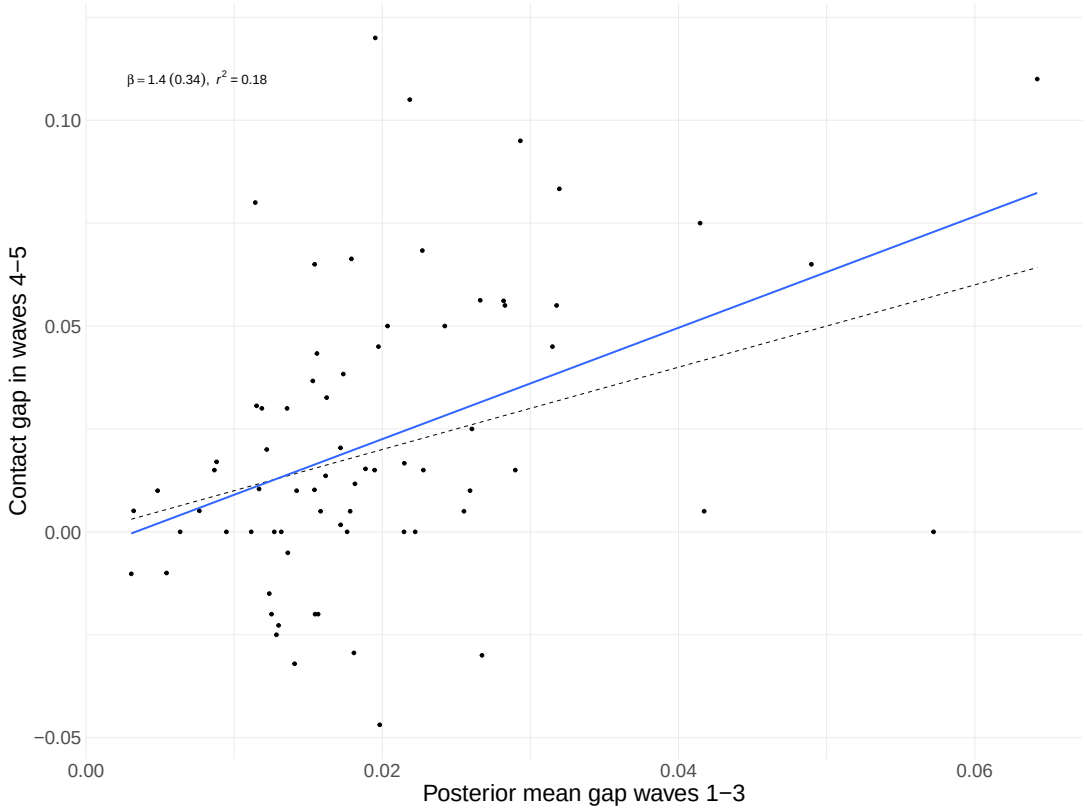
s of the distribution of firm-specific contact gaps for race and gender. The red solid line presents the empirical ashed line plots the CDF of population contact gaps based on the deconvolution estimates in Figure 6. The CDF of posterior means, constructed treating the deconvolved density as prior knowledge. The pink dashed nates shrunk linearly towards the grand mean with weights given by the signal-to-noise ratio $\hat{\theta}/(s_f^2+\hat{\theta})$, where e variance of contact gaps across firms.

Figure A12: Deconvolution of firm-level racial discrimination without support restriction



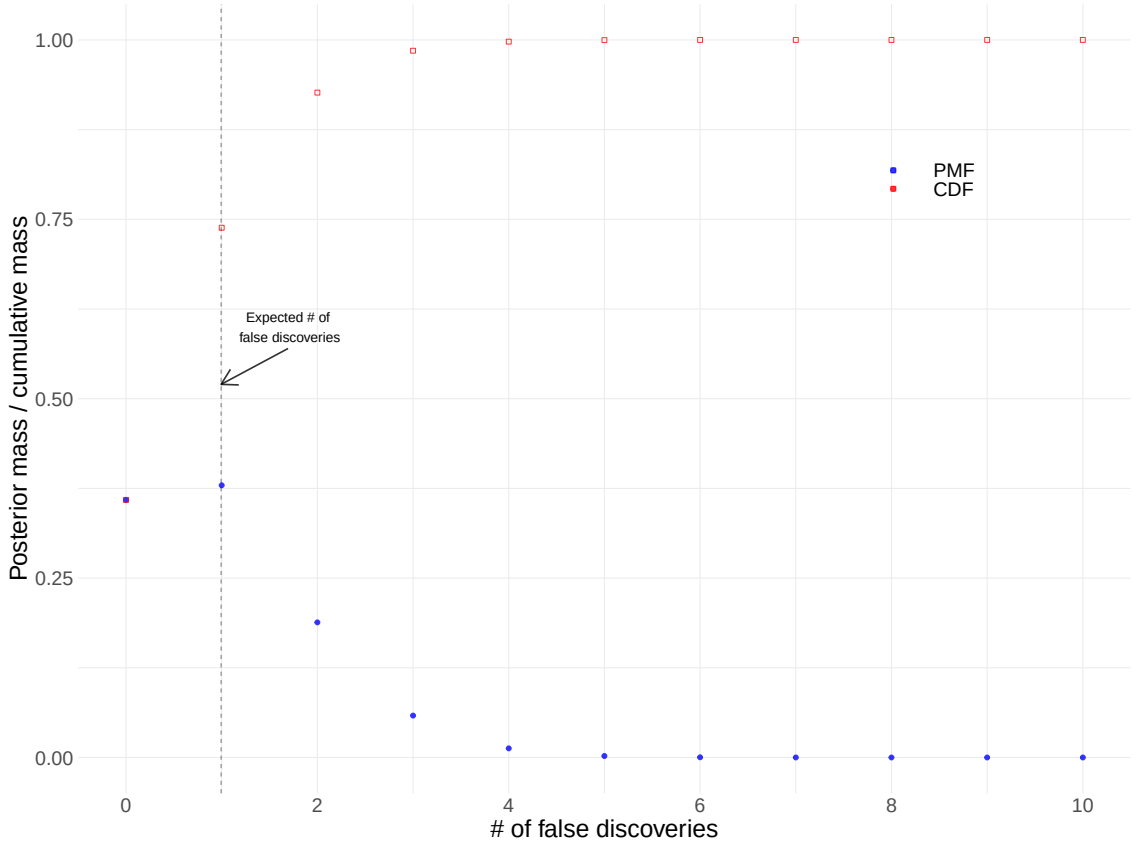
Notes: This figure presents non-parametric estimates of the distribution of firm-specific white-Black contact rate differences. The red histogram shows the distribution of estimated firm contact gaps. Blue line shows estimates of the population contact gap distributions. The population distributions are estimated by applying the deconvolveR package (Narasimhan and Efron, 2020) to firm-specific z-score estimates, then numerically integrating over the empirical distribution of standard errors 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 A13: Out-of-sample predictive power of racial contact gap posteriors



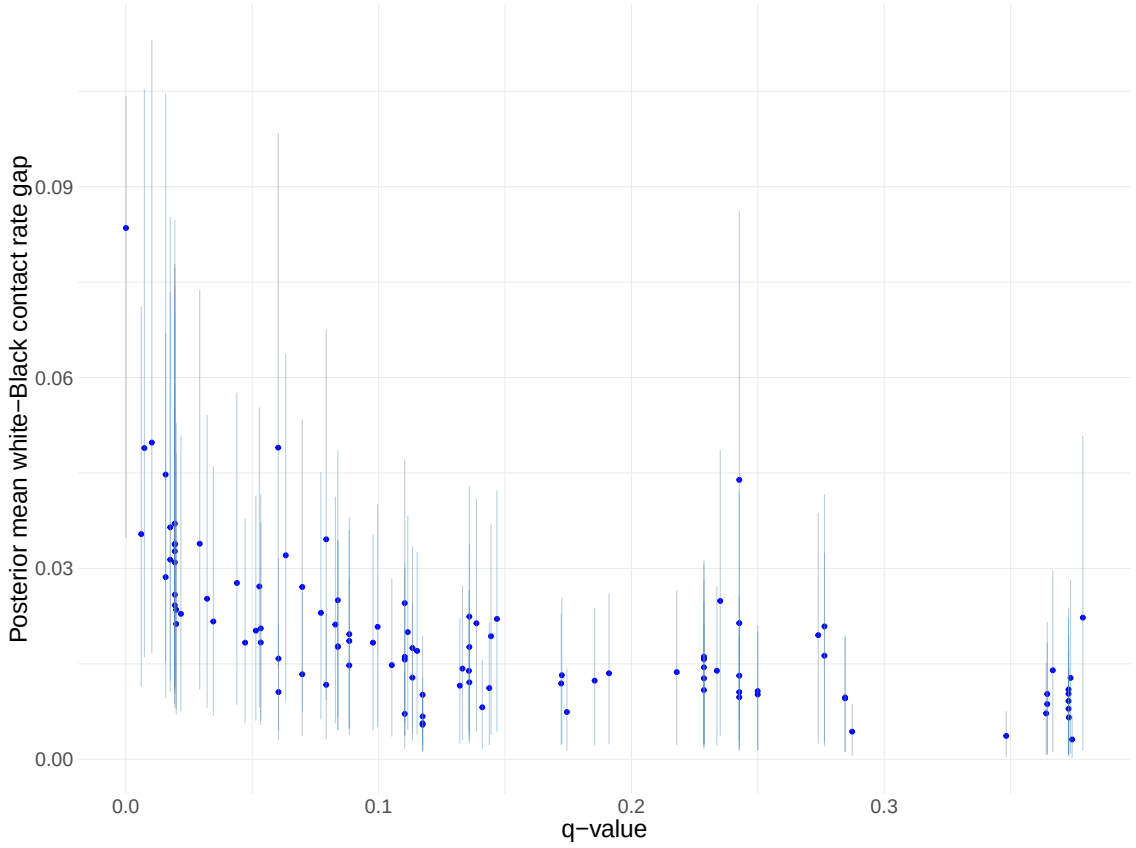
Notes: This figure plots posterior mean white-Black contact gaps computed using data from waves 1-3 against observed gaps in waves 4-5 for the sample of firms included in all five waves. Posterior means are computed using the population contact gap distributions estimated by applying the deconvolveR package (Narasimhan and Efron, 2020) to firm-specific z-score estimates from waves 1-3. The penalization parameter in the deconvolution step is calibrated so that the resulting distribution matches the corresponding bias-corrected variance estimate. The black dotted line is the 45 degree line. The blue line is the least squares fit. Adjusting for the noise in the wave 4 and 5 estimates yields a bias corrected R^2 of 0.5, or a correlation between predictions in later waves and the latent true contact gaps of $\sqrt{0.5} = 0.71$.

Figure A14: Posterior false discovery distribution among 23 firms with low q-values



Notes: This figure plots EB posterior estimates of the probability mass function and cumulative distribution of false discoveries among the 23 firms with q-values below 0.05 for race. Posterior was calculated using the Poisson binomial distribution implied by the 23 firms' LFDR estimates plotted in Figure 10. The dotted line denotes the expected number of false discoveries among these firms, which coincides with the mean of their LFDRs.

Figure A15: 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) χ^2 test of heterogeneity	(2) p-value for no discrim against:	(3) Bias-corrected	(4) Cross-wave	(5) Cross-state
Race	229.5	W: 1.00	0.0184	0.0171	0.0182
	[0.000]	B: 0.00	(0.0029)	(0.0032)	(0.0031)
Gender	124.2	M: 0.06	0.0207	0.0213	0.0200
	[0.000]	F: 0.03	(0.0044)	(0.0043)	(0.0045)
Over 40	90.2	Y: 0.15	0.0098	0.0096	0.0099
	[0.072]	O: 0.02	(0.0060)	(0.0067)	(0.0057)

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 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 2 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). Column 3 reports bias-corrected estimates of standard deviations of average contact gaps across firms based on covariances between job-level contact gaps. 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 the Appendix. Standard errors for all variance estimators are produced by job-clustered weighted bootstrap.

Table A2: Variance components for other resume attributes

			Contact gap SD		
	(1) χ^2 test of heterogeneity	(2) p-value for no discrim against:	(3) Bias-corrected	(4) Cross-wave	(5) Cross-state
LGBTQ Club Member	88.0	W: 1.00	-	-	-
	[0.885]	B: 0.98			
Gender Neutral Pronouns	126.5	Y: 0.92	0.0198	0.0177	0.0147
	[0.076]	O: 0.65	(0.0156)	(0.0176)	(0.0208)

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 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 2 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). Column 3 reports bias-corrected estimates of standard deviations of average contact gaps across firms based on covariances between job-level contact gaps. 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 the Appendix. Standard errors for all variance estimators are produced by job-clustered weighted bootstrap. 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	0.070	0.090	0.026
	(0.000)	(0.000)	(0.000)
Prevalence bound	0.073	0.000	0.014
	(0.012)	(0.001)	(0.021)

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 viols/emp: Includes all wage and hour compliance violations 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 adj wage, white - Black adj wage, male adj wage, male - female adj wage: Constructed using the CPS Monthly Outgoing Rotation Groups from 2009 to 2019,

Table C1: Summary statistics for firm-level covariates

	Mean	SD	Median
Firm performance			
Log employment	11.067	1.01	10.922
Sales / emp ($M)	0.331	0.36	0.238
Profit / emp ($M)	0.101	0.08	0.078
GD score	3.566	0.32	3.600
Legal compliance			
DOL viols / emp	0.136	0.37	0.002
Empl-discr cases / thousand emp	0.048	0.13	0.020
Federal contractor	0.667	0.47	1.000
Firm diversity			
% board Black	0.088	0.07	0.091
% board female	0.257	0.10	0.255
% managers non-white	0.257	0.09	0.250
% managers female	0.493	0.39	0.449
Has chief diversity officer	0.167	0.37	0.000
GD diversity score	3.816	0.33	3.800
Callback patterns			
Callback centralization	-1.117	0.38	-1.073
Observations	108		

Notes: This table reports summary statistics for firm-level covariates. See Appendix C for full details on the sources and construction of each variable.

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 correspond to the same coefficients from regressions with an indicator for female (or Black when constructing gender gaps), education (6 categories), and a quartic in age also included. 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.

Table C2: Firm-level predictors of centralization

Firm performance	
Log employment	-0.144
	(0.116)
Sales / emp	-0.0682
	(0.0813)
Profit / emp	-0.00380
	(0.0797)
GD score	-0.211
	(0.173)
Legal compliance	
DOL viols / emp	-0.0836
	(0.121)
Empl-discr cases / emp	0.0576
	(0.0427)
Federal contractor	0.559**
	(0.265)
Firm diversity	
% board Black	0.178
	(0.116)
% board female	-0.0153
	(0.0982)
% managers non-white	0.00869
	(0.123)
% managers female	0.0553
	(0.0986)
Has chief diversity officer	0.148
	(0.211)
GD diversity score	0.223
	(0.157)
Observations	10500

Notes: This table reports the multivariate relationship between centralization and other firm-level predictors. All predictors except the binary indicators for federal contractor status and having a chief diversity officer are normalized to have standard deviation of 1. As with firm-level relationships reported in Figure 5, the regression is estimated on job-level data with firm-clustered standard errors. See Appendix C for full details on the sources and construction of each variable.

- 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").
  - 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

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

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, 1988; Aaronson, Barrow and Sander, 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 $(\tau_1,\tau_2) \in \{1,...,T_f\}^2$

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

where $\hat{\bar{\Delta}}_{f\tau}$ 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

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

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:

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

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:

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

Appendix E Alternative Deconvolution Estimates

This section explores the robustness of estimated population contact gap distributions to alternative models for the relationship between estimated gaps, $\hat{\Delta}_f$, and their standard errors, s_f. Our baseline analysis assumes that s_f is independent of the population z-score Δ_f/s_f. After applying the Efron (2016) deconvolution estimator to the sample z-scores $\hat{\Delta}_f/s_f$, we recover the distribution of Δ_f by numerically integrating against the empirical distribution of standard errors. Here we consider three alternatives: a variance stabilizating transformation approach, a "local deconvolution" approach that separately estimates population distributions among groups of firms with similar standard errors, and a non-parametric approach that estimates the joint distribution of contact gaps and standard errors.

Appendix E.1 Variance stabilizing transformation

If one assumes a parametric model for the dependence of the firm specific variances on the latent contact gaps of the form $s_f^2 = h(\Delta_f)$, then heteroscedasticity can be eliminated via the variance stabilizing transformation:

$$y(t) = \int_{\infty}^{t} h(x)^{-1/2} dx.$$

Note that $\frac{d}{dt}y(t) = h(t)^{-1/2}$. Hence, standard delta-method reasoning implies that $y(\hat{\Delta}_f)|\Delta_f \sim \mathcal{N}(y(\Delta_f), 1)$. Applying the deconvolution estimator of Efron (2016) to the transformed estimates $y(\hat{\Delta}_f)$, one can then generate an estimate of the population distribution of Δ_f using the change of variables $\hat{g}_\Delta(x) = \hat{g}_{y(\Delta)}(x)h'(x)$, where $\hat{g}_t(\cdot)$ is the estimated density of t.

To implement this approach, we allow for non-linear dependence of the (squared) standard errors on contact gaps by assuming that

$$h(\Delta) = \alpha + \beta_1\Delta + \beta_2\Delta^2 \quad \text{for } \Delta \in \mathcal{S},$$

where $\mathcal{S}$ is the support of the population contact gap under study. We use split-sample IV (Angrist and Krueger, 1995) to estimate the parameters (β_1, β_2). Splitting each firm's jobs into two groups $g \in \{0, 1\}$, we proxy each group's values of Δ_f and Δ_f^2 with $\hat{\Delta}_{fg}$ and $\hat{\Delta}_{fg}^2 - s_{fg}^2$, respectively, where s_{fg} is the standard error of $\hat{\Delta}_{fg}$. We then estimate β_1 and β_2 via a regression of s_{fg}^2 on $(\hat{\Delta}_{fg}, \hat{\Delta}_{fg}^2 - s_{fg}^2)$ using as instruments $(\hat{\Delta}_{f(g)}, \hat{\Delta}_{f(g)}^2 - s_{f(g)}^2)$, where $(g) = 1 - g$ refers to the omitted group. To minimize uncertainty attributable to the splitting process, we take the median across 1,000 iterations of this procedure.

Figure E1 presents the resulting deconvolved population distributions of contact gaps for race and gender. As in the primary estimates, the race gap distribution exhibits a

peak close to zero and a fat right tail of heavy discriminators. The distribution of gender gaps continues to show concentrated mass near zero and severe discriminators in both tails. Figure E2 summarizes the concentration of discrimination based on the variance-stabilized approach by plotting the Lorenz curves implied by the deconvolved density $\hat{g}_\Delta$. Similar to our baseline estimates in Figure 7, these curves imply discrimination is concentrated in a relatively small share of firms for both race and gender. Finally, Figure E3 shows the estimated racial contact gap distribution based on the variance-stabilization approach without restricting the density of Δ_f to be weakly positive, which produces minimal changes to the results.

Appendix E.2 Local deconvolutions

A less parametric approach to dealing with heteroscedasticity in the contact gaps is to simply estimate the deconvolution within bins defined by ranges of s_f. This approach weakens the requirement that s_f be independent of the population z-score Δ_f/s_f in the full population to a requirement that independence only hold among firms with similar s_f. To implement this approach, we split firms into two groups $k \in \{H, L\}$ by whether their contact gap standard error falls above / below the sample median standard error. We then apply the deconvolution estimator of Efron (2016) to the sample z-scores in each group, $\hat{\Delta}_{fk}/s_{fk}$, and recover the group-specific population contact gap density, $g_{\Delta,k}$, by integrating against the empirical distribution of standard errors in that group. The marginal density of contact gaps is then given by the mixture:

$$\hat{g}_\Delta(x) = \frac{1}{2} g_{\Delta,H} + \frac{1}{2} g_{\Delta,L}$$

We use a common penalization parameter in the deconvolution step for both groups and calibrate it so that the resulting marginal distribution matches the corresponding bias-corrected variance estimate from Table 4.

Figure E4 shows the resulting group-specific densities for both race and gender. Figure E5 shows the corresponding marginal densities. As in the primary estimates, the race density shows concentrated mass close to zero and fat right tail. The gender density is strongly peaked at zero. Both densities continue to show that discrimination is strongly concentrated in a relatively small share of firms, as shown in Lorenz curves presented in Figure E6.

The close agreement of the top 20% share estimates found in Figures 7, E6, and E2 is reassuring and suggests our modeling of heteroscedasticity patterns is not an important factor driving our concentration results.

Appendix E.3 NPMLE

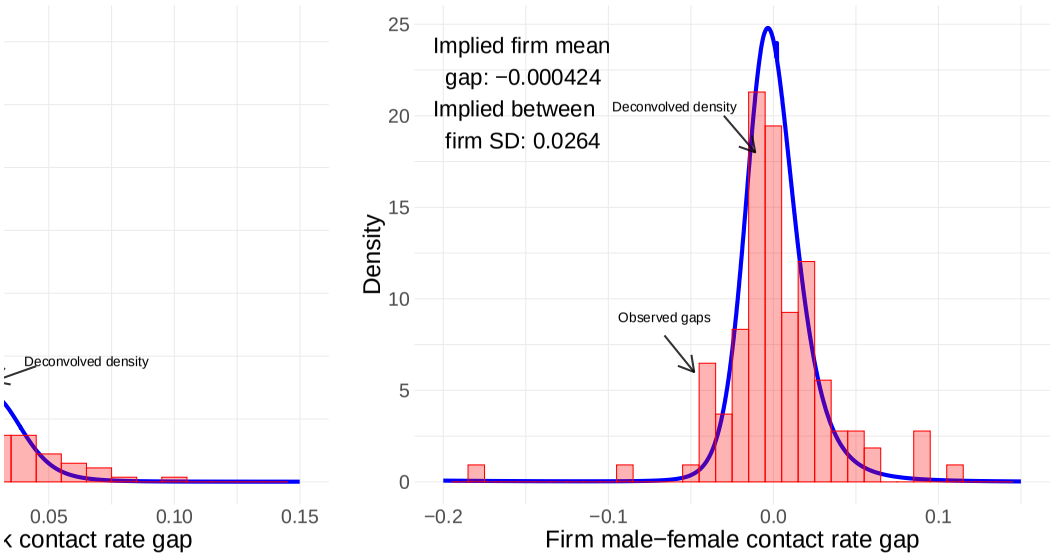
As a final approach to accounting for heteroscedasticity, we estimate a non-parametric mixing distribution over contact gaps and standard errors that allows for unrestricted dependence between these objects. To implement this approach, we use the approximation to the Kiefer-Wolfowitz non-parametric maximum likelihood estimator (NPMLE) developed by Koenker and Mizera (2014) and implemented in the REBayes package of Koenker and Gu (2017).

Figure E7 presents the resulting discrete marginal densities of contact gaps for race and gender. NPMLE estimates of the distribution of racial discrimination show similar patterns to our earlier spline approximations, with a concentrated mass of firms exhibiting limited discrimination and a fat tail of more heavy discriminators. Gender estimates show substantial mass near zero and smaller mass points in the extremes of both tails.

Figure E8 shows that these densities also imply substantial concentration of discrimination among a subset of employers. For comparability with prior results, we linearly interpolate between mass points, which yields kinks in the Lorenz curves. The interpolated top 20% shares are slightly higher than in our baseline specification utilizing spline approximations, suggesting again that our concentration finding is highly robust. The Gini coefficients are also close to those found in our baseline specification, with the race Gini slightly higher and the gender Gini slightly lower than the corresponding estimates in Figure 7.

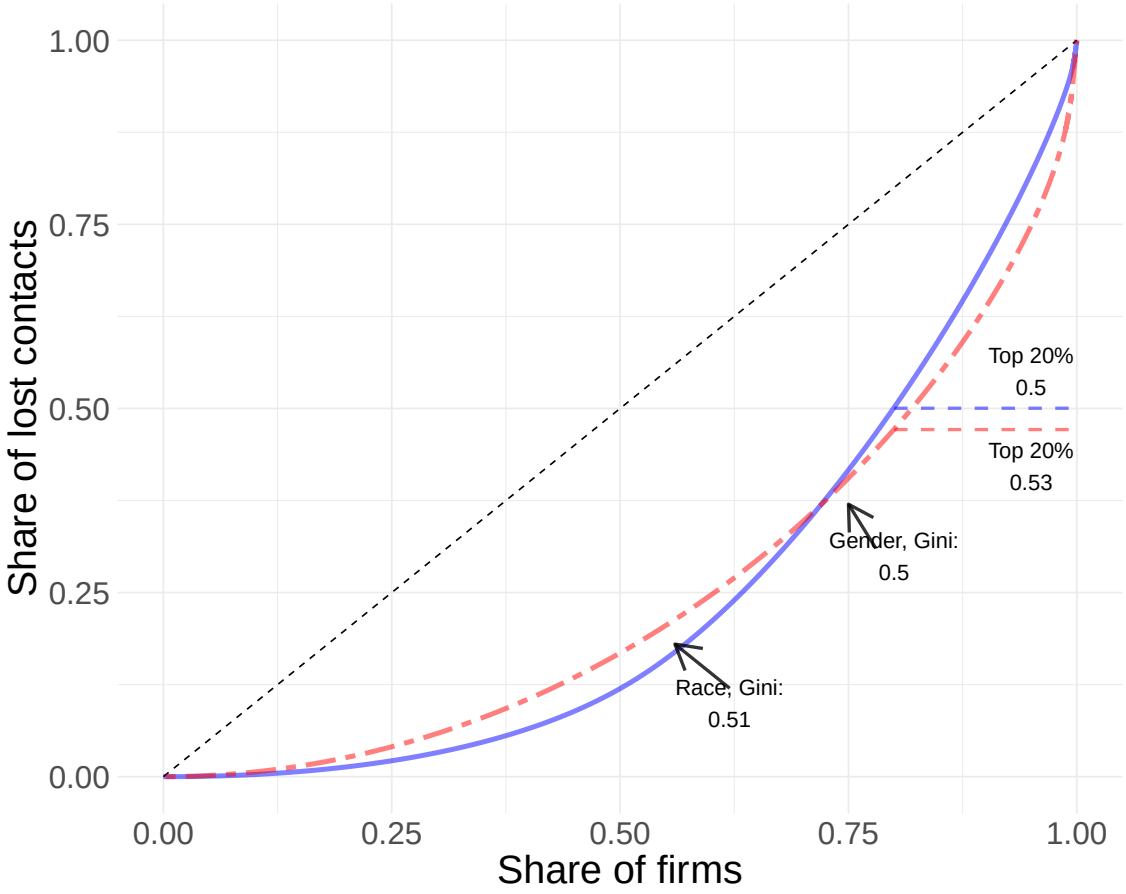
1: Variance-stabilized deconvolution of firm-level discrimination distributions

ce b) Gender



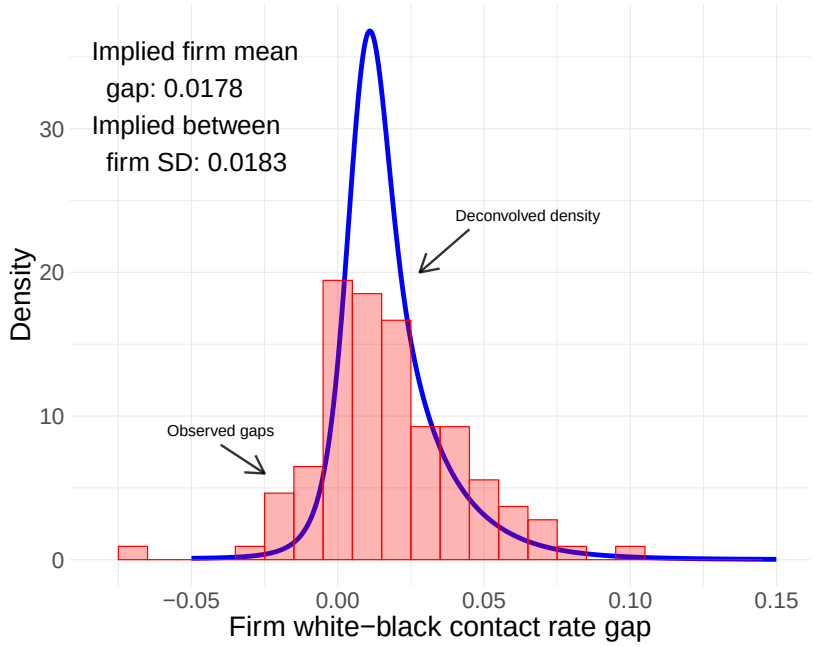
metric estimates of the distribution of firm-specific contact gaps. Panel (a) presents estimates for white-Black) presents estimates for male-female differences. Red histograms show the distribution of estimated firm contact opulation contact gap distributions. The population distributions are estimated by applying the deconvolveR 20) to variance-stabilized estimates of firm-specific contact gaps. The variance-stabilizing transformation is $+\beta_1\Delta_f+\beta_2\Delta_f^2$, with α, β_1, and β_2 estimated via split-sample IV (Angrist and Krueger, 1995). The estimated ed gaps is transformed into the distribution of Δ_f using the change of variables formula. The penalization is calibrated so that the resulting distribution matches the corresponding bias-corrected variance estimate from population Δ_f is constrained to be weakly positive.

Figure E2: Variance-stabilized 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 E1. 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.

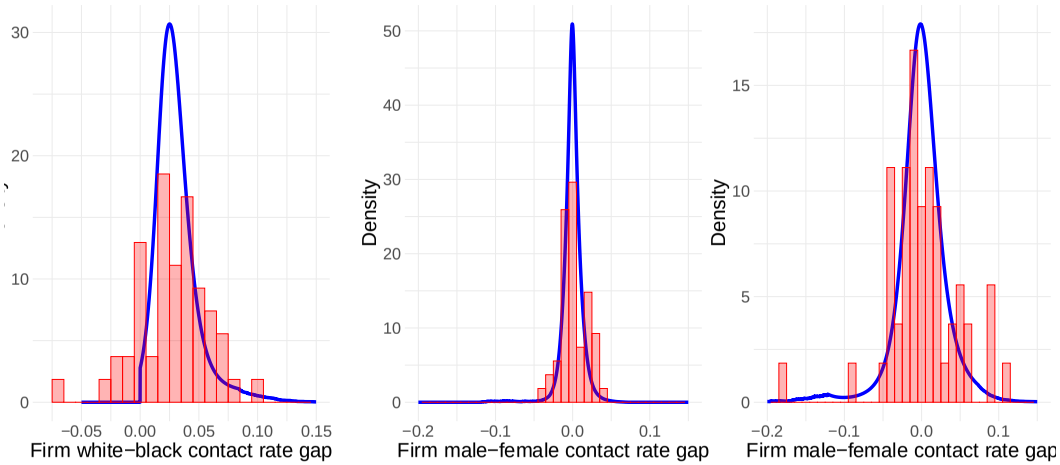
Figure E3: Variance-stabilized deconvolutions of racial discrimination without support restriction



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 an estimate of the population contact gap distribution constructed as in Panel (a) of Figure E1, but without the restriction that the density of Δ_f is weakly positive.

gure E4: Local deconvolutions of firm-level discrimination distributions

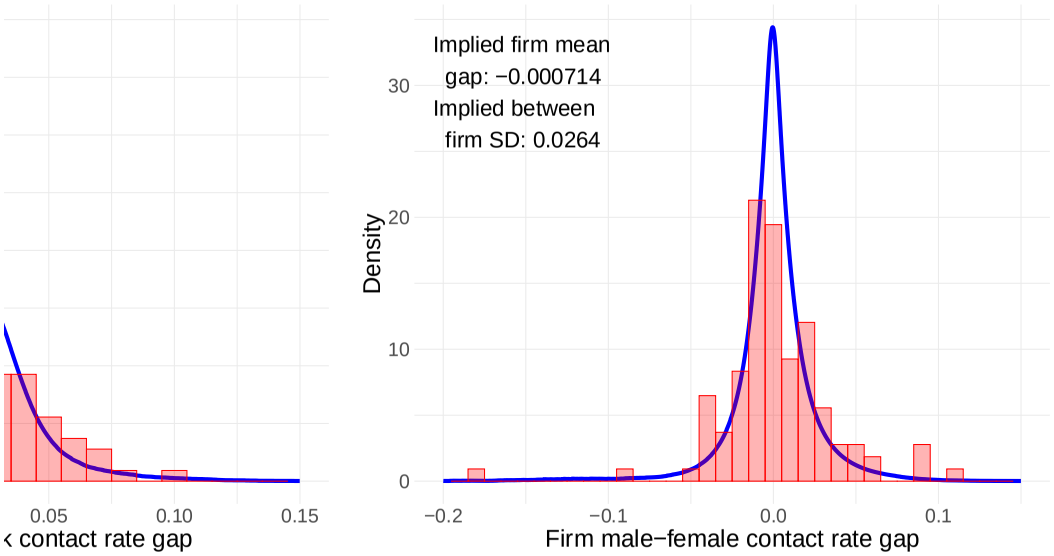
ce b) Gender



umetric estimates of the distribution of firm-specific contact gaps estimated separately for firms with above / (a) presents estimates for white-Black contact rate differences, and panel (b) presents estimates for male-female e distribution of estimated firm contact gaps in each group. Blue lines show estimates of population contact e population distributions are estimated by applying the deconvolveR package (Narasimhan and Efron, 2020) hin group, then numerically integrating over the group's empirical distribution of standard errors. A common e deconvolution step for both groups and calibrated so that the resulting marginal distribution matches the e estimate from Table 4. In panel (a), the density of population z-scores is constrained to be weakly positive

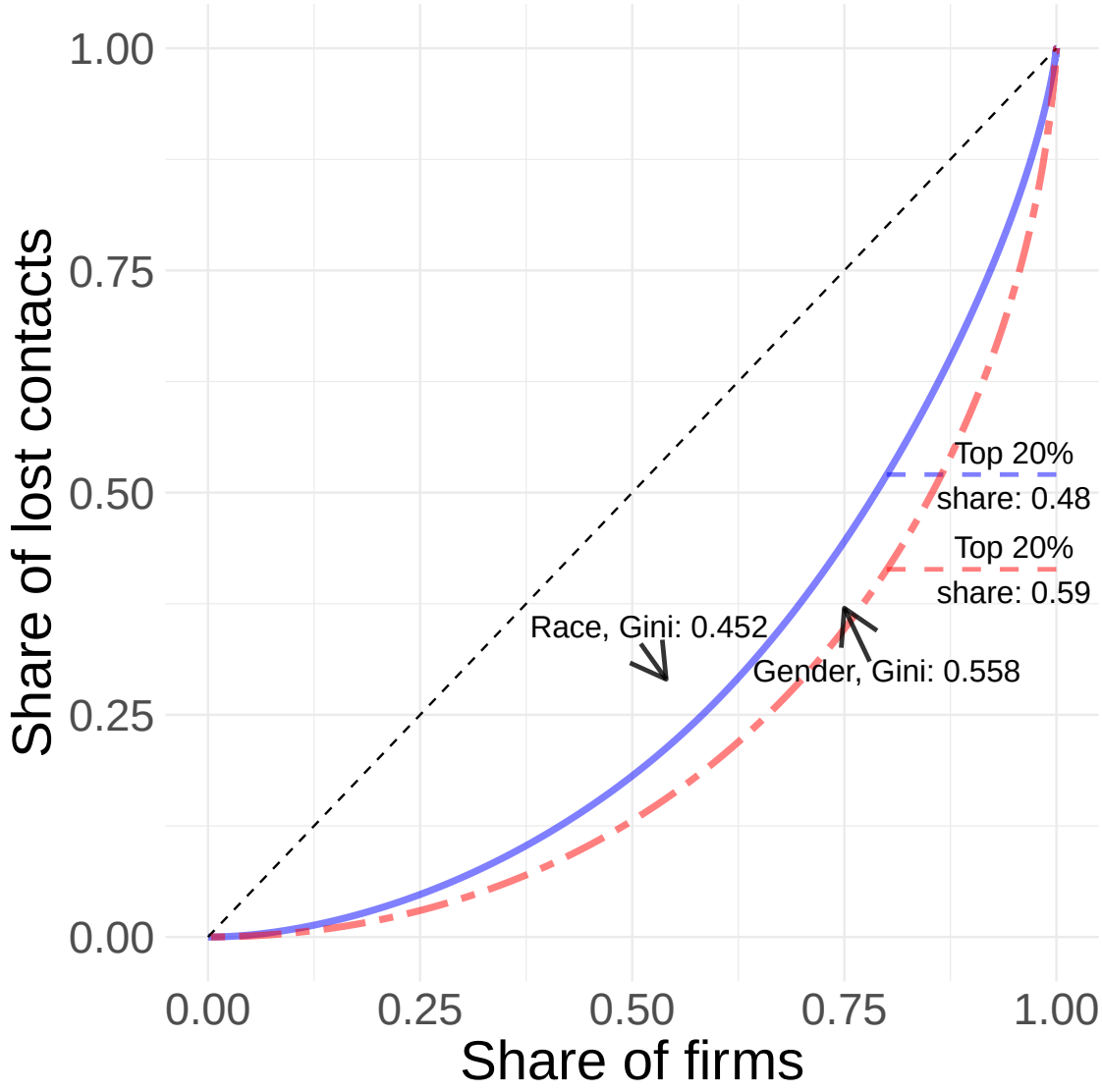
E5: Marginal distributions of firm-level discrimination from local approach

ce b) Gender



metric estimates of the marginal distribution of firm-specific contact gaps corresponding to the group-specific presents estimates for white-Black contact rate differences, and panel (b) presents estimates for male-female distribution of estimated firm contact gaps. Blue lines shows estimates of population contact gap distributions. the average of the group-specific densities in Figure E4.

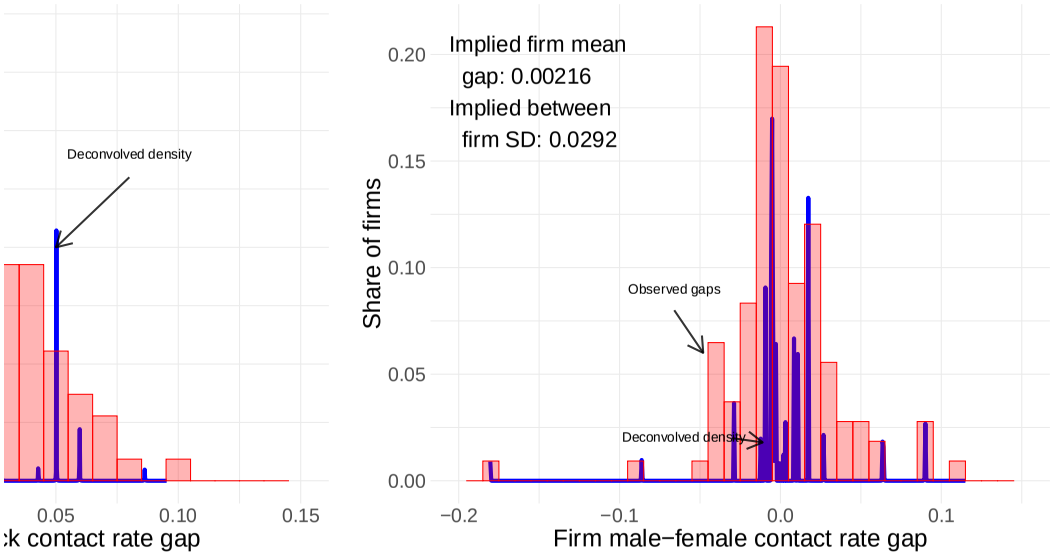
Figure E6: Local deconvolution Lorenz curves



Notes: This figure displays Lorenz curves implied by the non-parametric deconvolution estimates of race and gender contact gap distributions in Figure E5. 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.

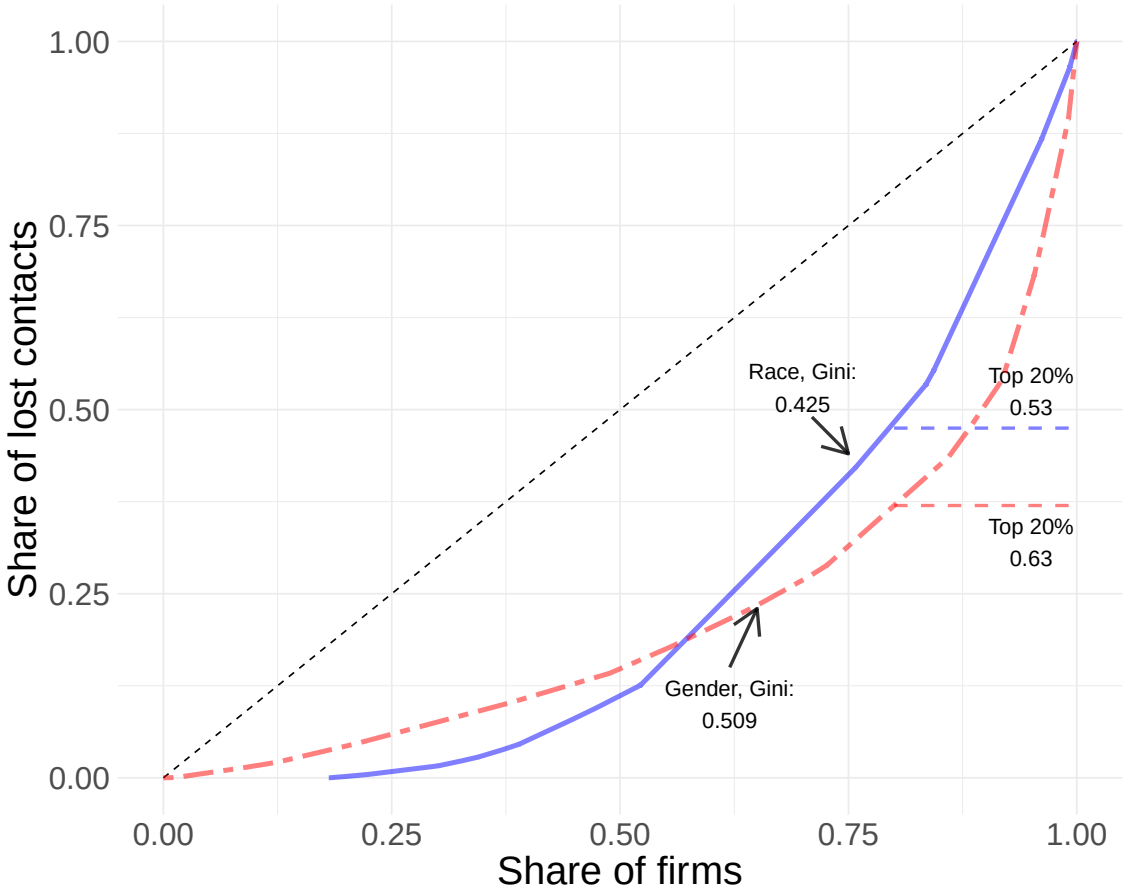
E7: NPMLE estimates of marginal distributions of firm-level discrimination

ce b) Gender



ametric maximum likelihood estimates of the distribution of firm-specific contact gaps estimated using the Panel (a) presents estimates for white-Black contact rate differences, where we impose the restriction that all l panel (b) presents estimates for male-female differences. Red histograms show the distribution of estimated estimates of population contact gap distributions. Population distributions are estimated allowing a non- the mixing distribution of contact gaps and standard errors. The figures plot the marginal distribution of is discrete, the blue lines plot the probability mass function in below, while the histogram reports the share of

Figure E8: NPMLE Lorenz curves



Notes: This figure displays Lorenz curves implied by the NPMLE estimates of race and gender contact gap distributions reported in Figure E7. 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. Linear interpolation has been used between mass points, which generates kinks in the curve. 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.

E-lit-008: The State of Hiring Discrimination: A Meta-Analysis of (Almost) All Recent Correspondence Experiments

Louis Lippens, Siel Vermeiren and Stijn Baert | European Economic Review 151: 104315 | 2023

https://cris.vub.be/ws/files/96282081/Lippens_et_al._2022_EER_The_state_of_hiring_discrimination.pdf

About this source

Lippens, Vermeiren and Baert build the largest register of correspondence experiments assembled to date and pool it with random-effects meta-analysis. On race, ethnicity and national origin the pooled discrimination ratio is 0.7113 across 143 experiments and 340,262 applications: minority-signaled applicants receive roughly 29 percent fewer positive employer responses. The penalty varies sharply by group — 0.5937 for Arab, Maghrebi or Middle Eastern applicants, 0.6845 for African, African American or Black applicants, and a statistically insignificant 0.9220 for Hispanic, Latin American or Caribbean applicants — and by region, with the Americas at 0.8027 the mildest of the large regions and Europe at 0.7004. That Americas figure matters directly for the section's hypothesis: a ratio of 0.80 means white-named applicants get about 25 percent more callbacks, roughly half the 50 percent advantage Bertrand and Mullainathan reported, so the canonical magnitude is not the pooled modern one. On time, the paper cuts both ways. Pooled across all regions there is a small but significant narrowing (coefficient 0.0173 on the year the experiment ended, p = 0.030; fitted ratio rising from 0.6053 in 2006 to 0.7605 in 2020), but that narrowing is driven entirely by Europe (r = -0.37, k = 94), it collapses to insignificance once the authors control for which minority groups each experiment studied or run country-by-country trends, and in the Americas there is no significant trend at all (r = 0.13, k = 38, p = 0.422) — a footnote adds that if anything the American trend points toward more discrimination. The authors' own summary is that hiring discrimination remains a pervasive issue with little change over time. They also show that discrimination against applicants with disabilities (0.5885), older applicants (0.6867) and less physically attractive applicants (0.6308) is at least as severe as the ethnic penalty, and that how a study defines a callback (interview invitation versus any positive reply) does not significantly change the measured gap.

Bound as the original PDF. Retrieved 2026-09-10. sha256 72ac6ebd7acef2c8...

Bound into an Economic Briefing · by Claude, may contain mistakes

Full document begins on the next page.

E-lit-008: https://cris.vub.be/ws/files/96282081/Lippens_et_al._2022_EER_The_state_of_hiring_discrimination.pdf

Vrije Universiteit Brussel

VRIJE UNIVERSITEIT BRUSSEL

The state of hiring discrimination: A meta-analysis of (almost) all recent correspondence experiments

Lippens, Louis; Vermeiren, Siel; Baert, Stijn

Published in:
European Economic Review

DOI:
10.1016/j.euroecorev.2022.104315

Publication date:
2023

License:
CC BY-NC-ND

Document Version:
Final published version

Link to publication

Citation for published version (APA):
Lippens, L., Vermeiren, S., & Baert, S. (2023). The state of hiring discrimination: A meta-analysis of (almost) all recent correspondence experiments. *European Economic Review*, *151*, Article 104315. https://doi.org/10.1016/j.euroecorev.2022.104315

Copyright
No part of this publication may be reproduced or transmitted in any form, without the prior written permission of the author(s) or other rights holders to whom publication rights have been transferred, unless permitted by a license attached to the publication (a Creative Commons license or other), or unless exceptions to copyright law apply.

Take down policy
If you believe that this document infringes your copyright or other rights, please contact openaccess@vub.be, with details of the nature of the infringement. We will investigate the claim and if justified, we will take the appropriate steps.

Download date: 10. Sep. 2026

European Economic Review 151 (2023) 104315



Contents lists available at ScienceDirect

European Economic Review

journal homepage: www.elsevier.com/locate/eer





The state of hiring discrimination: A meta-analysis of (almost) all recent correspondence experiments

Louis Lippens [a,b,*], Siel Vermeiren [a], Stijn Baert [a,c]

[a] Ghent University
[b] Vrije Universiteit Brussel
[c] University of Antwerp; Université catholique de Louvain; Institute of Labor Economics (IZA); Global Labor Organization (GLO)

ARTICLE INFO

JEL classification:
J71
J23
J14
J15
J16

Keywords:
Hiring discrimination
Unequal treatment
Meta-analysis
Correspondence experiment
Audit study

ABSTRACT

Notwithstanding the improved integration of various minority groups in the workforce, unequal treatment in hiring still hinders many individuals' access to the labour market. To tackle this inaccessibility, it is essential to know which and to what extent minority groups face hiring discrimination. This meta-analysis synthesises a quasi-exhaustive register of correspondence experiments on hiring discrimination published between 2005 and 2020. Using a random-effects model, we computed pooled discrimination ratios concerning ten discrimination grounds upon which unequal treatment in hiring is forbidden by law. Our meta-analysis shows that hiring discrimination against candidates with disabilities, older candidates, and less physically attractive candidates seems equally severe as the unequal treatment of candidates with salient racial or ethnic characteristics. Moreover, hiring discrimination against older applicants is more prominent in Europe than in the United States. Last, while we initially find a significant decrease in ethnic hiring discrimination in (Western) Europe, we find no structural evidence of recent temporal changes in hiring discrimination when controlling for the minority groups considered, at the country level, or based on the various other grounds within the scope of this review.

1. Introduction

Although the workforce has become increasingly diverse—improving the integration of female, migrant, and older workers, amongst other groups—many individuals belonging to various minority groups still face considerable discrimination in the labour market (Organisation for Economic Co-operation and Development [OECD], 2020a). In part because of their decreased chances for labour market access, these individuals are at elevated risk of long-term unemployment and labour market inactivity (OECD, 2020a, 2020b). This underutilisation of talent could result in needless economic costs for firms and society (Baert, 2021; OECD, 2020a; Pager, 2016). For policymakers, it is vital to know which (minority) groups are confronted with hiring discrimination and to understand the severity of this labour market's inaccessibility. In this way, targeted diversity policies, such as outreach campaigns focusing on underrepresented or discriminated groups, can be implemented to help those who require said policies the most (OECD, 2020a).

Research on labour market discrimination has long focused on the non-experimental decomposition approach to measure discrimination (Blinder, 1973; Kitagawa, 1955; Neumark, 2018; Oaxaca, 1973). This approach has historically involved isolating the

* Corresponding author at: Louis Lippens, Faculty of Economics and Business Administration, Ghent University, Sint-Pietersplein 6, 9000 Ghent, Belgium; louis.lippens@ugent.be
E-mail address: louis.lippens@ugent.be (L. Lippens).

https://doi.org/10.1016/j.euroecorev.2022.104315
Received 17 January 2022; Received in revised form 31 August 2022; Accepted 2 October 2022
Available online 20 October 2022
0014-2921/© 2022 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

impact of discrimination on wages via regression analyses (Borjas, 2020). Variance that could not be explained by differences in human capital between groups of interest (e.g. Blacks and Whites) was consequently attributed to discrimination. However, it is difficult to capture the true amount of variance explained by human capital under this approach, primarily due to omitted variable bias (Altonji and Blank, 1999; Borjas, 2020).[1] The decomposition method thus sketches an incomplete picture of discrimination (Borjas, 2020; Gaddis, 2018).

To overcome the limitation of the decomposition approach, researchers began to use audit studies as an alternative experimental method to measure the incidence of labour market discrimination (Gaddis, 2018). At first, this was mainly done by sending out pairs of real applicants (i.e. actors) who differed in terms of visible characteristics based on which unequal treatment is forbidden (e.g. skin colour) to interview for the same job. Differences in job offers were subsequently interpreted in terms of discrimination. In the early 2000s, however, Bertrand and Mullainathan (2004) steered the research area of labour market discrimination in a different direction: correspondence audits replaced in-person audits as the standard for measuring hiring discrimination (Gaddis, 2018).[2] Rather than sending out actors as applicants, these correspondence experiments consisted of mailing written applications from fictitious job seekers in response to real job postings. By randomly assigning individual characteristics based on which selection is forbidden, the effect of these characteristics on employers' reactions can be given causal interpretations. Compared with in-person audits, the perceived differences between applicants, produced by minute differences in their behaviour during the interview, are nullified. Moreover, the application process is less resource-intensive. Because of its high employability and the causal interpretation that underpins its results, the correspondence testing method is still considered the reference method for measuring hiring discrimination at present (Baert, 2018; Neumark, 2018; Verhaeghe, 2022).

In recent years, a considerable number of scholars have reviewed and synthesised parts of the hiring discrimination literature, focusing on the correspondence testing method. We know of nineteen topical meta-studies in this evolving research area: Adamovic (2020, 2022), Baert (2018); Bartkoski et al. (2018); Batinović et al. (2022); Bertrand and Duflo (2017); Derous and Ryan (2019), Flage, (2020), Gaddis et al. (2021), Heath and Di Stasio (2019), Lippens et al. (2022), Neumark (2018), Quillian and Midtbøen (2021), Quillian et al. (2017, 2019, 2020), Rich (2014), Thijssen et al. (2021), and Zschirnt and Ruedin (2016). Table A1 in Appendix A provides an overview of the aforementioned meta-studies including details about the (i) type of review, (ii) type of analysis, (iii) inclusion of a meta-regression component, (iv) included discrimination grounds, (v) type of studies considered, (vi) region, (vii) period, and (viii) main findings.[3] These studies can be roughly classified into two categories: traditional reviews and systematic reviews (Briner and Denyer, 2012). Traditional reviews, on the one hand, typically lack structure in the scoping or search process; it is not always clear why some studies are highlighted while others are ignored. Few recent meta-studies can be categorised as such. Systematic reviews, on the other hand, have a clear scope and target specific studies meeting predetermined inclusion criteria. The majority of recent meta-studies summarising correspondence audits are systematic reviews of which most also pursue meta-analytic methods to quantify and contextualise hiring discrimination.

All of these meta-studies have merit. Several of these studies have provided insightful empirical, theoretical or policy-orientated observations but, for understandable reasons, have only focused on specific grounds of discrimination (predominantly race, ethnicity, and national origin) while neglecting others (e.g. Bartkoski et al., 2018; Gaddis et al., 2021; Quillian et al., 2017). Other reviews taking a broader view of hiring discrimination have brought forth equally interesting insights but do not provide a systematic account of the existing correspondence experiment literature (e.g. Bertrand and Duflo, 2017; Neumark, 2018; Rich, 2014). Moreover, recent meta-studies have successfully implemented meta-regression techniques to identify interesting patterns in the data. For example, Gaddis et al. (2021) found that controlling for relevant covariates, discrimination against Black Americans is highest in high-stake settings such as hiring and housing (versus education, medical services, or public services). Quillian et al. (2019) found that there is less discrimination in jobs where a college degree is required (versus a high school degree or equivalent). Yet, again, the scope of these studies has been limited to race, ethnicity, and national origin. Baert (2018) was the first in attempting to counter these limitations by (i) adopting a broad view on hiring discrimination considering all grounds based on which unequal treatment is forbidden under United States federal and state law and (ii) providing a quasi-exhaustive register of correspondence experiments conducted since Bertrand and Mullainathan's (2004) seminal study. However, the main limitation of Baert's (2018) work is the absence of meta-analysis to synthesise and compare the results of the included studies.

The current study documents the most extensive register of correspondence experiments on hiring discrimination to date. We compile and synthesise a comprehensive catalogue of correspondence experiments published between 2005 and 2020, predominantly from the Americas, Europe, and Asia. We simultaneously provide an overview of the presently under-researched grounds for discrimination and the hiatuses that still exist in the current literature on hiring discrimination. Altogether, we gather 306 correspondence experiments (i.e. units of observation in our analysis) originating from 169 separate correspondence audit studies. Adding up the job applications from each correspondence experiment, these experiments comprise almost 965,000 fictitious applications in response to real job vacancies.

[1] Classic examples of omitted variables include (but are not limited to) unobserved supply-side factors such as personal motivation or choices as well as possessing an extensive professional network.

[2] Bertrand and Mullainathan (2004) were not the first to implement the correspondence audit method (see e.g. Jowell & Prescott-Clarke, 1970). However, their influential study gave traction to the so-called 'third wave' (and the current fourth wave) of audit studies which mainly comprised correspondence audits (Gaddis, 2018). Since then, the publication density of correspondence audits has increased substantially.

[3] Other meta-studies that do not report on audits studies related to hiring discrimination were not considered for this overview. One example thereof is Drydakis (2022), who recently examined the literature on the relationship between sexual orientation and earnings differences.

Our synthesis offers scholars and policymakers an understanding of the prevalence and severity of hiring discrimination concerning various discrimination grounds within the scope of this review. Specifically, we meta-analytically quantify hiring discrimination regarding ten discrimination grounds upon which unequal treatment is forbidden under United States federal or state law, including (i) race, ethnicity, and national origin, (ii) gender and motherhood status, (iii) age, (iv) religion, (v) disability, (vi) sexual orientation, (vii) physical appearance, (viii) wealth, (ix) marital status, and (x) military service or affiliation. The standardised meta-analytical approach enables comparisons of levels of hiring discrimination across discrimination grounds and minority groups.[4] We also assess heterogeneity in hiring discrimination, providing a more granular perspective of our findings. For each discrimination ground, we explore whether (i) levels of hiring discrimination are related to how call-backs are reported and measured, (ii) persistent regional differences in hiring discrimination exist, or (iii) unequal treatment in hiring has changed in recent times. In summary, this is the first study using meta-analytic techniques to make such a broad comparison of previous experiments on hiring discrimination, across discrimination grounds and minority groups, based on the currently largest documented set of correspondence audit studies from across the globe.

2. Data and methods

In this section, we elaborate on (i) the scope of our meta-analysis; (ii) how we identified and selected studies; (iii) which variables we collected from these studies, as well as how we classified some of them into broader categories to identify differences across these categories; and (iv) the details of the meta-analytical methods we used to analyse the resulting data. In this process, we paid special attention to the reporting guidelines for meta-analyses in economics of Havránek et al. (2020)—we refer the reader to Appendix A (Table A2) for the corresponding checklist.

2.1. Scope

We used various eligibility criteria based on the Population, Intervention, Comparison, Outcome (PICO) framework to delineate our review (Richardson et al., 1995).[5] Table 1 provides an overview of these criteria. We limited our review to correspondence studies in which unequal treatment in hiring was assessed between fictitious applicants belonging to minority groups and their majority counterparts. We considered studies written in English that were first published as a discussion paper, pre-print, or journal article between 2005 (the year after Bertrand and Mullainathan's seminal 2004 correspondence study) and 2020 (the most recent full calendar year at the time this study was conducted) in particular.[6]

Similar to the delineation of the discrimination grounds in Baert's (2018) correspondence experiment register, we limited our scope to the forms of hiring discrimination prohibited under United States federal or state law.[7] We focused on United States federal and state law for two reasons. First, we wanted to ensure the complementarity of our dataset with Baert (2018). Their original reasoning was to consider legal discrimination grounds where most correspondence audit studies (focusing on hiring discrimination) were conducted—this was (and still is) the United States. Second, compared to the European Union, where many audit studies originate from, there is less room for the discretionary application of employment discrimination law compared to the United States (Ganty and Benito Sanchez, 2021). We thus took into account the following discrimination grounds: (i) race, ethnicity, and national origin, (ii) gender and motherhood status, (iii) religion, (iv) disability, (v) age, (vi) military service or affiliation, (vii) wealth, (viii) genetic information, (ix) citizenship status, (x) marital status, (xi) sexual orientation, (xii) political affiliation, (xiii) union affiliation, and (xiv) physical appearance. In our final analysis, we retained only ten of these grounds because, similar to Baert (2018), (i) no (new) correspondence experiments related to genetic information or citizenship status were identified in the search process; and (ii) we found only one experiment related to political orientation and one experiment related to union affiliation, from which we could not calculate pooled discrimination ratios.

2.2. Study selection

We used multiple sources to identify, screen, and select eligible studies for our meta-analysis. Fig. 1 depicts a structured overview of this process. First, we identified potentially eligible correspondence studies. On the one hand, we sourced studies included in Baert's (2018) register of correspondence experiments, which resulted from an elaborate systematic search for correspondence experiments on hiring discrimination. On the other hand, we performed a systematic search on the Web of Science and Google Scholar databases in spring 2021. Our search used the keywords 'correspondence experiment', 'correspondence study', 'fictitious resume', 'fictitious cv',

[4] We do not attempt to provide a reasoning for the underlying mechanisms of the uncovered hiring discrimination. Our primary goal remained to compare levels of hiring discrimination across discrimination grounds and minority groups. For recent overview studies on the empirical evidence concerning the economic mechanisms of (ethnic) labour market discrimination, we refer the reader to Lang & Kahn-Lang Spitzer (2020), Lippens et al. (2022), and Neumark (2018).

[5] We extended the PICO framework to be more specific in the delineation of the scope of our review. Precisely, we also considered 'study type', 'context', and 'timing' and excluded 'intervention' because it was not relevant to our search query.

[6] The specific year allocated to a given study was based on the year the study was initially published. For example, Larsen and Di Stasio (2021) was first published online as a pre-print in 2019 before appearing officially in the *Journal of Ethnic and Migration Studies* in 2021.

[7] A reassessment of United States federal and state law was made in late 2020. Relative to Baert (2018), not much had changed. One noteworthy change, however, is that discrimination based on LGBT+ status has been made illegal at the US federal level (Boystock v. Clayton County, 2020).

Table 1
Eligibility criteria for study inclusion.

Criterion	Details
Study type	Correspondence experiment in which applications were sent in response to vacancies.
Population	(Fictitious) applicants from various minority groups and their majority counterparts.
Outcome	Disadvantageous, unequal treatment in the hiring and selection process (i.e. hiring discrimination).
Comparison	Hiring chances of minority applicants compared with those of majority applicants.
Context	Hiring discrimination related to the grounds upon which unequal treatment is forbidden under United States federal or state law (i.e. race, ethnicity, and national origin, gender and motherhood status, religion, disability, age, military service or affiliation, wealth, genetic information, citizenship status, marital status, sexual orientation, political orientation, union affiliation, and physical appearance).
Timing	Study first published between 2005 and 2020.

Notes. The framework used to define the eligibility criteria is based on the PICO (Population, Intervention, Comparison, Outcome) framework first coined by Richardson et al. (1995).

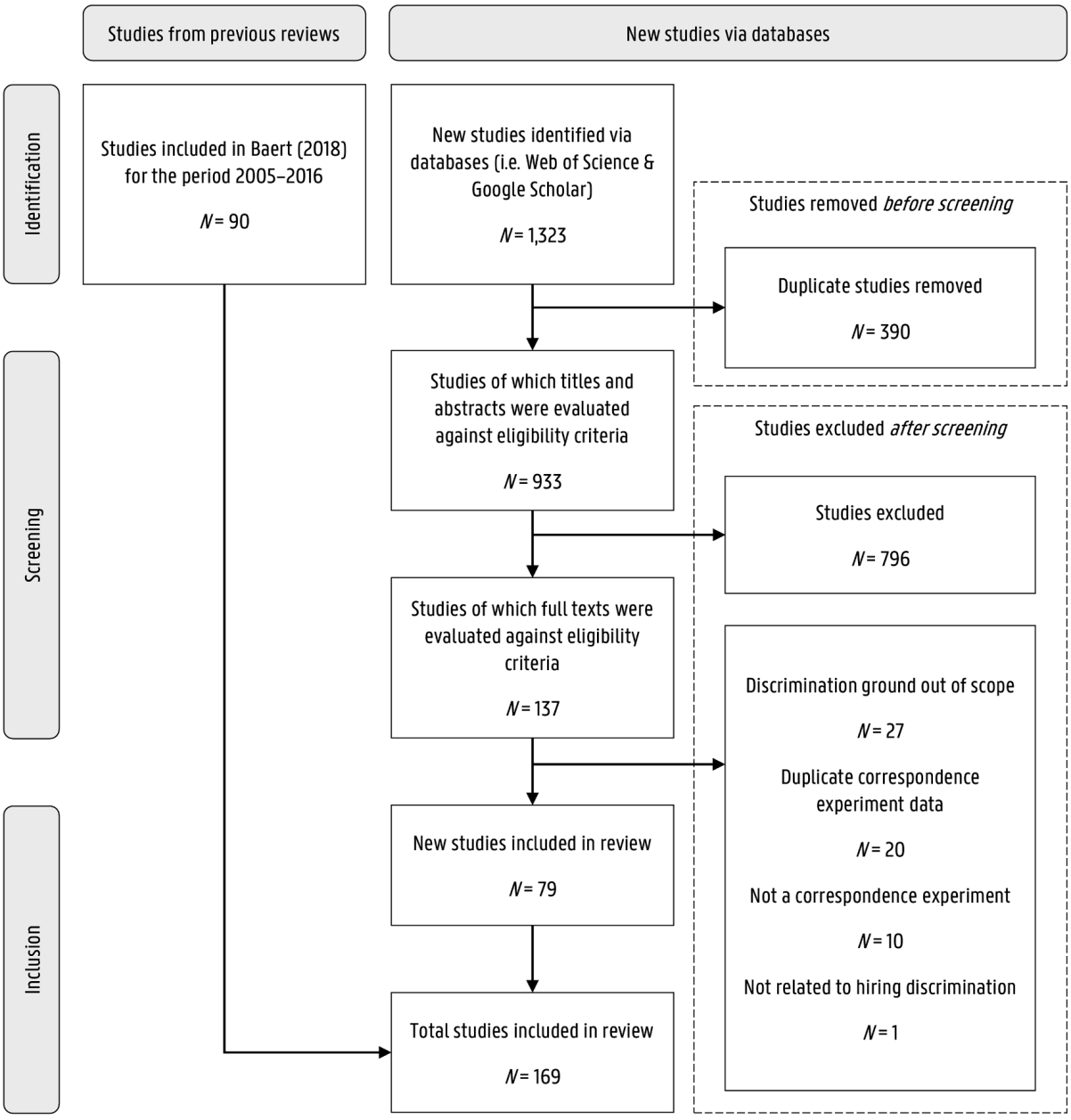


Fig. 1. Study selection flow diagram. Notes. This figure is adapted from Page et al., 5).

'fictitious application', and 'field experiment' in combination with the keyword 'discrimination'. In general, we confined our search to studies published in the period 2005 to 2020. To extend this systematic search, we also performed a cited reference search with the references from Baert's (2018) book chapter as the input of our queries

Next, we appraised the studies that had not already been identified by Baert (2018). In total, we evaluated the titles and abstracts of 933 studies against our eligibility criteria (see Section 2.1). After an initial screening of the titles and abstracts, we reviewed the full text of the remaining 137 articles. The risk of reviewer bias was reduced by having two researchers independently review the selected articles. After this review, 79 studies were identified that fully matched the criteria. There were four reasons for excluding certain studies after appraising their full text: (i) unequal treatment based on the discrimination ground in the scope of the study was not forbidden under United States law ($N = 27$, 46.55% of the total number of excluded full texts);[8] (ii) the correspondence experiment was entirely based on data used in a previously published (and already included) study ($N = 20$, 34.48%);[9] (iii) the study did not use the correspondence testing method (e.g. in-person audit; $N = 10$, 17.24%); or (iv) the study was solely related to housing discrimination instead of hiring discrimination ($N = 1$, 1.72%).

We retained a total of 169 studies, of which 90 were already included in Baert's (2018) book chapter, resulting in 306 units of observation. There are more units of observation than studies due to our definition of a 'unit of observation', i.e. a unique correspondence experiment based on the related (i) discrimination ground, (ii) treatment group, (iii) control group, and (iv) region where the test was performed. For example, Di Stasio et al. (2021) considered hiring discrimination against Muslims in Germany, Norway, Spain, the Netherlands, and the United Kingdom. To allow for heterogeneity analyses on the basis of region (see Sections 2.4.2 and 3.4.2), this study was subdivided into multiple units of observation stemming from the same study.

2.3. Data collection

We captured a multitude of variables for each correspondence experiment. First, we registered the basic information of the studies, including the authors' names and the year the article was officially published. In addition to the latter, we also recorded (i) the year the study was initially published (e.g. as a pre-print or early-access article), which was the year we used when evaluating the article against our eligibility criteria, and (ii) the year the correspondence experiment ended.[10] Second, we documented where the research took place, including the country and (sub-)region. The latter was based on the M49 Standard for geographic regions of the United Nations (2021; see Table A3 for a tabulated overview). Third, we registered the (experimental) treatment group and the control group of the correspondence experiment. The specific treatment groups identified in the included studies were classified into broader groups to facilitate further analyses. Because no common global framework of ethnic and racial minority groups exists, the classification of these groups consisted of a proprietary framework based on how various governmental bodies of OECD member countries collect and categorise diversity data (Balestra and Fleischer, 2018; European Commission, 2021; Morning, 2008).[11] The classification related to the other discrimination grounds was based on the logical grouping of the respective treatment groups. The final classification can be found in Appendix A (Table A4).

Fourth, we documented data related to the outcomes of the correspondence experiments. We captured the overall treatment effect of the results in the original studies (averaged across sub-groups at the experiment level). We also recorded the classification of the outcome variable (i.e. call-back). If a call-back consisted of an invitation to a job interview (or any broadly defined positive response of the employer, such as a request for additional information), we labelled it narrow (or broad).

Most importantly, we registered the number of observations (i.e. fictitious job applications) and the number of positive call-backs in both the treatment and control groups. The accuracy of these variables was independently assessed and verified by at least two authors. Outcome measure data required to calculate pooled discrimination ratios were missing for 32 experiments (9.82%, out of a total of 326 units). After contacting the corresponding authors of the respective studies to retrieve these data, 12 cases could be completed (37.50% of cases with missing data), meaning that we had no data for the remaining 20 units of observation.[12] These units were excluded from the meta-analysis, resulting in 306 valid units (cfr. supra). Reporting bias, which could (partly) originate from missing data, was formally evaluated when we tested for publication bias (see Section 2.4.3).

From these data, we derived a standardised discrimination ratio. The specification of this ratio is shown in Eq. (1). The

[8] For example, Gaddis, (2015) looks at unequal treatment based on the educational institution an applicant attended. Making selection decisions based on the educational institution is not a ground for (illegal) discrimination. While the author makes an interesting assessment by looking at the interaction of this criterion with race, the results from this particular part of the study were not included in our analyses.

[9] Some studies we initially identified were based on data already contained in other studies. Including identical data multiple times in our analyses would obviously bias the pooled discrimination ratios. Simultaneously, this would lower the variance around these estimates. To avoid this 'multiple publication bias', we excluded such studies from our review (see Page et al., 2020).

[10] There is a discrepancy between the temporal period of this review (2005–2020) and the timeframe used in the heterogeneity analyses (see section 3.1 and 3.4.2). This is because the latter is based on the year in which the correspondence experiment ended. The rationale for this is that this time variable more accurately represents the timing of the experiment (vis-à-vis the year the research was published).

[11] In their correspondence test, Jacquement and Yannelis (2012), for example, assigned African American names to the minority group, while Gaddis, (2015) used Black-sounding names. In the United States and the United Kingdom, these origins are both classified as 'African (American)' or 'Black'. Therefore, we created the category 'African/African American/Black' as an umbrella term for similar treatment groups.

[12] These missing data were linked to ten studies: Beam et al. (2020), Carlsson and Eriksson (2019), Darolia et al. (2016), Drydakis (2017), Guul et al. (2019),Patacchini et al. (2015), Stone and Wright (2013), Thijssen, Coenders, & Lancee (2021), Thomas (2018), and Yemane and Fernández-Reino (2021).

discrimination ratio is a risk ratio (or relative risk) equal to the division of two proportions: (i) the proportion of positive call-backs in the treatment group (a_k) relative to the total number of observations in that group ($n_{k\ treat}$), and (ii) the proportion of positive call-backs in the control group (c_k) relative to the total number of observations in that group ($n_{k\ control}$). Because the discrimination ratio can be interpreted in terms of relative change, a ratio of 0.75, for example, indicates a 25% reduction in positive call-backs of the (fictitious) applicants of the minority group vis-à-vis the applicants of the control group, aggregated at the level of the correspondence experiment. Since our estimation strategy assumed that the included discrimination ratios follow a normal-like distribution, we log-transformed these ratios before pooling them in our meta-analysis (see Section 2.4.1). This approach ensured that opposite, same-sized effects were equidistant. In addition, Eq. (2) illustrates the calculation of the standard error of these log-transformed discrimination ratios (also see Harrer et al., 2021).

$$DR_k = \frac{a_k / n_{k\ treat}}{c_k / n_{k\ control}} \tag{1}$$

$$SE_{\ln DR} = \sqrt{\frac{1}{a_k} + \frac{1}{c_k} - \frac{1}{n_{k\ treat}} - \frac{1}{n_{k\ control}}} \tag{2}$$

2.4. Analyses

Our synthesis was based on the results of the correspondence experiments identified and selected in the previous steps. Our goals were (i) to quantify and compare the level of hiring discrimination for each of the various discrimination grounds and treatment groups in the scope of our analysis and (ii) to identify heterogeneity in hiring discrimination based on (a) the definition of the call-back variable, (b) the region where the correspondence experiment took place, and (c) the period of the experiment. We used R (version 4.1.0) for our analyses and relied on the {meta} package for most of our calculations (e.g. estimating the pooled ratios or detecting reporting bias; Balduzzi et al., 2019). We also used the {dmetar} package to identify influential cases, the {metasens} package to perform 'limit' meta-analyses, and the {metafor} package to perform meta-regression analyses and to examine the statistical (in) dependence of the sampled discrimination ratios (Harrer et al., 2019; Schwarzer et al., 2020; Viechtbauer, 2010).

2.4.1. Pooled discrimination ratios

To quantify the level of hiring discrimination across the various discrimination grounds, we used a random-effects model to pool the discrimination ratios of the included studies by discrimination ground and treatment group. We opted for this model because it starts from the premise that the true level of discrimination varies across studies. We assumed that there was at least some variation in these levels caused by (subtle) differences in the (i) definition and conceptualisation of the treatment and control groups, (ii) measurement of the responses (or call-backs), and (iii) overall experimental design and process. We applied Knapp–Hartung adjustments when calculating the confidence intervals around the pooled discrimination ratio estimates (Knapp and Hartung, 2003). This method assumes a *t*-distribution of the pooled effect rather than a normal distribution, which reduces the chance of obtaining false-positive results (Langan et al., 2019). The Knapp–Hartung adjustments produce more conservative (i.e. wider) confidence interval estimates than when these adjustments would not be applied.

To calculate the weights of the studies (w) in the reported pooled discrimination ratios, we used the commonly reported Mantel–Haenszel method for binary outcome data—the formula is shown in Eq. (3) (for more details, see Borenstein et al., 2009; Mantel and Haenszel, 1959). This method takes into account the number of cases in the treatment and control groups wherein the call-back was positive (a and c, respectively), as well as the number of cases in the treatment and control groups wherein the call-back was negative or absent (b and d, respectively; Mantel and Haenszel, 1959). This approach inherently attaches more importance to studies with larger sample sizes or overall higher numbers of positive call-backs. To generate more balanced weights, the weights were adjusted for between-study variance (τ^2). This procedure decreases potential overemphasis (or underemphasis) on studies with a relatively large (or small) sample size (see also Section 2.4.2; Borenstein et al., 2009). Subsequently, these variance-adjusted weights (w^*) were plugged into the general specification of the random-effects model, as illustrated in Eq. (4). Here, $\widehat{DR}$ is the pooled discrimination ratio, DR_k represents the observed discrimination ratio of the individual correspondence experiments, $\widehat{\zeta}$ is the error related to the overarching distribution of true discrimination ratios, and $\widehat{\varepsilon}$ symbolises the sampling error (Borenstein et al., 2009; Harrer et al., 2021).

$$w_k = \frac{(a_k + b_k)^* c_k}{a_k + b_k + c_k + d_k} \tag{3}$$

$$\widehat{DR} = \frac{\sum_{k=1}^{K} (DR_k + \widehat{\zeta}_k + \widehat{\varepsilon}_k)^* w_k^*}{\sum_{k=1}^{K} w_k^*} \tag{4}$$

2.4.2. Heterogeneity analyses

To meaningfully interpret the pooled discrimination ratios by discrimination ground and treatment group and to identify differences in hiring discrimination levels, we quantified and examined variability in statistical and design-related heterogeneity. First, we

assessed statistical heterogeneity by calculating a statistic that captured the variability in the true discrimination ratios underlying the data (Rücker et al., 2008). More specifically, we calculated I^2 estimates, which indicate the proportion of the total variability due to between-study variability (i.e. a value between 0 and 1) in the true discrimination ratios not caused by sampling error (Harrer et al., 2021; Higgins and Thompson, 2002; Veroniki et al., 2015). The I^2 statistic compares the studies' discrimination ratios to the pooled ratio, weighted by the inverse of the variance of the respective studies, taking into account the total number of studies. Therefore, this statistic is insensitive to (substantial) changes in the number of studies included in the analysis (Cochran, 1954; Harrer et al., 2021; Hoaglin, 2016). A high I^2 value (closer to 1) warrants exploring the source of this heterogeneity, for example by investigating moderation effects, as well as checking and controlling for possible extreme cases (i.e. outliers) in the dataset.[13]

Second, we evaluated design-related heterogeneity (i.e. heterogeneity due to differing designs across studies) by performing meta-regression analyses using the weighted least squares method (WLS) with a maximum-likelihood estimator. More specifically, we examined the heterogeneity in hiring discrimination for the following study-level variables: (i) the call-back classification, (ii) the geographical area where the correspondence experiment took place, and (iii) the year when the experiment ended. We did so largely in an exploratory manner—we tested how hiring discrimination contextually varied and compared our results with those of previous meta-studies where relevant. This approach contributed to (partly) explaining the statistical heterogeneity estimated in the previous step. Following Schwarzer et al., (2015) guidelines, we only performed meta-regressions on the groups of studies for which the total number of included studies was equal to or greater than ten.

The general meta-regression specification is given in Eq. (5), analogous to the notation in Harrer et al. (2021). In this equation, $\widehat{DR_k}$ is the observed discrimination ratio of each correspondence experiment k, $\widehat{DR}$ is the estimated pooled discrimination ratio, $\widehat{\beta}$ is the coefficient representing the fixed effect, x is the study-level variable, p stands for the number of predictors, $\widehat{\varepsilon}$ symbolises the sampling error, and $\widehat{\zeta}$ is the error related to the overarching distribution of true discrimination ratios representing the random effect (also see Section 2.4.1). In the reported models, standard errors related to the model coefficients were clustered at the study level (Viechtbauer, 2010).

$$\widehat{DR_k} = \widehat{DR} + \widehat{\beta}_1 x_{1k} + \ldots + \widehat{\beta}_p x_{pk} + \widehat{\varepsilon}_k + \widehat{\zeta}_k \tag{5}$$

2.4.3. Publication bias

There are several reasons why publication bias could adversely impact the results from our meta-analysis, resulting in either an over- or underestimation of hiring discrimination in certain cases. For example, some studies could have been withheld from publication because the results were non-significant, uninteresting or inconclusive while other studies with interesting, significant, or substantial effects were not (i.e. outcome reporting bias). There is also a risk of language bias as only studies in English were included in the review (i.e. language bias). Furthermore, the time lag between the conducting of the research and the publishing of the research could make some studies stay under the radar (i.e. time-lag bias; see Section 3.1).

To measure and counter the impact of publication bias, we applied three analytic techniques: (i) a graphical inspection of funnel plot asymmetry, (ii) the calculation of a 'bias statistic' of funnel plot asymmetry, and (iii) the calculation of bias-adjusted hiring discrimination estimates through 'limit' meta-analyses (Harrer et al., 2021). First of all, we constructed contour-enhanced funnel plots setting off the discrimination ratio of the correspondence experiments against their standard errors. These plots were overlaid by multiple funnel shapes depicting (i) the 95% and 99% confidence intervals around the estimated pooled discrimination ratio for a given discrimination ground or treatment group wherein the observations are expected to fall and (ii) the 90%, 95%, and 99% confidence intervals around the null effect (or discrimination ratio of 1) for a given discrimination ground or treatment group (see Figure A2–1 to Figure A2–16). The second set of funnel shapes helped us in distinguishing outcome reporting bias from other types of publication bias because we could evaluate whether there might be an underreporting of null results.

Second, we used Peters' (2006) binary-effects adaptation of Egger's regression test to calculate a 'bias statistic' for assessing funnel plot asymmetry, which formally compares the discrimination ratios of the respective studies against their standard errors—its null hypothesis assumes that there is no asymmetry. In line with Harrer et al. (2021) and Sterne et al. (2011), we only calculated bias statistics if the total number of correspondence audits for a given analysis equalled or exceeded ten, otherwise, the statistical power could be too low to detect asymmetry.

Third, we replaced some of the original estimates with bias-adjusted estimates obtained from 'limit' meta-analyses. Through these analyses, we allowed for interactions between the observed effects, on the one hand, and the standard error of the pooled effect and the between-study variance, on the other hand (Rücker et al., 2011). This analytical approach resulted in so-called 'shrunken' discrimination ratios that largely account for small-study publication bias (Harrer et al., 2021; Schwarzer et al., 2020). Smaller-sized correspondence experiments are, on average, at greater risk of only being reported if they produce large, statistically significant effects vis-à-vis experiments with larger sample sizes (Borenstein et al., 2009; Harrer et al., 2021). Small-study effects can thus be a source of publication bias due to the correlation between a study's publication status and the nature of its findings. As a general rule, we only

[13] Higgins and Thompson's (2002) guidelines state that values around 25%, 50%, and 75% indicate low, moderate, and high heterogeneity, respectively.

calculated these bias-adjusted discrimination ratios (and replaced the original estimates with these ratios) if the total number of correspondence audits for a given analysis equalled or exceeded ten (Harrer et al., 2021; Schwarzer et al., 2015).[14] If not, of the subset of meta-analyses containing fewer than ten studies, the statistical heterogeneity might be too high and the number of observations too low to meaningfully interpret the bias-adjusted discrimination ratios. In cases where we did not calculate bias-adjusted discrimination ratios, we reported the original estimates.[15] The statistical significance of the differences between the original discrimination ratios and the robust, bias-adjusted versions of the discrimination ratios was assessed using *z*-tests (see Table A11; for details on the computational approach, see Altman & Bland, 2003).

2.4.4. Robustness analyses

To further assess the robustness of our results, we (i) measured and controlled for outliers (i.e. influential cases that substantially affect the pooled discrimination ratio), (ii) evaluated potential statistical dependence between the sampled discrimination ratios, and (iii) recalculated *p*-values of the meta-regression coefficients via a permutation test (Borenstein et al., 2009; Higgins et al., 2019; Viechtbauer et al., 2015). First, we identified outliers by looking at studies with extremely small and large discrimination ratios. This process is relevant because it gives us an indication of whether removing these extreme cases from the analyses results in distinctly different discrimination ratios and thus how robust the original estimates are controlling for these outliers. Pooled ratios that are based on only a few correspondence experiments (i.e. units of observation) might be particularly impacted by outliers. In line with Harrer et al. (2021), we defined said ratios as those for which the upper (lower) bound of the 95% confidence interval was lower (higher) than the lower (upper) bound of the confidence interval of the pooled discrimination ratio. To clarify, this means that the ratios of these influential cases were so extreme that they significantly differed from the pooled ratio (at the 5% significance level). Eventually, we recalculated the pooled discrimination ratios, excluding these outliers, and evaluated whether they significantly differed from the original estimates.

Second, we examined the statistical independence of the sampled discrimination ratios. Interdependency between the discrimination ratios could arise in cases wherein different ratios relied on observations from the same control group (Higgins et al., 2019). For example, if a given experiment consisted of an unmatched design with two distinct treatment groups A and B and one control group C, both $\widehat{DR}_{A-C}$ (i.e. the discrimination ratio comparing A with C) and $\widehat{DR}_{B-C}$ (i.e. the discrimination ratio comparing B with C) would be partly based on identical information related to the same control group. This factor could lead to the underestimation of between-study variability, which could, in turn, result in false-positive pooled discrimination ratios. To examine potential statistical independence, we fitted three-level mixed models including estimates of between-study and within-study heterogeneity as well as two-level models that only included estimates of within-study heterogeneity per treatment group and compared these models using ANOVA (for the computational approach, see Harrer et al., 2021). We found no evidence that the three-level models had a better fit with the data than the two-level models (see Table A10). We can thus assume that our results were not significantly impacted by interdependency between the sampled discrimination ratios.

Third, on each estimated meta-regression model, we performed a permutation test with 1000 iterations to control for possible model overfitting (Harrer et al., 2021; Viechtbauer et al., 2015). This test boils down to iteratively redrawing samples from the underlying data, reordering the data into different permutations, to recalculate the *p*-values associated with the test statistics related to the model coefficients. In essence, this test enabled us to better evaluate whether the coefficients identify true patterns in the data or whether they model statistical noise (Harrer et al., 2021). The output from these permutation tests did not significantly alter the interpretation of our meta-regression results.

3. Results

In Section 3.1, we provide some descriptive statistics regarding the correspondence experiments included in our meta-analysis. Subsequently, in Sections 3.2–3.4, we concentrate on the meta-analytic statistics: (i) the pooled discrimination ratios by discrimination ground, (ii) the heterogeneity of these ratios by treatment group, and (iii) their heterogeneity by call-back classification, region, and period. Where appropriate and relevant, the statistical heterogeneity (i.e. statistical measures quantifying between-study variability) and the robustness of the results are discussed. In Section 3.5, finally, we comment on the potential impact of publication bias. The quasi-exhaustive register of correspondence experiments published between 2005 and 2020 on which our analyses were based can be retrieved in full from Table B1 in Appendix B.

3.1. Descriptive statistics

Fig. 2 shows an increase in the annual number of studies based on the correspondence testing method published between 2005 and 2020. More specifically, the number of ended experiments rises as of 2005—right after the publication of Bertrand and Mullainathan's (2004) study—and continues to increase steadily in subsequent years. There is a remarkable peak in the number of publications in

[14] In general, we want to caution the reader when interpreting the pooled estimates for which the included number of correspondence experiments is lower than ten. The uncertainty around the pooled estimates based on a small number of studies can become large, especially in cases where the estimate is based on very few experiments (e.g. only two or three).

[15] In previous versions of this paper (e.g. the IZA discussion paper; Lippens et al., 2021), the bias-adjusted estimates were only reported in the appendix and the reporting of the results was entirely based on the unadjusted estimates.

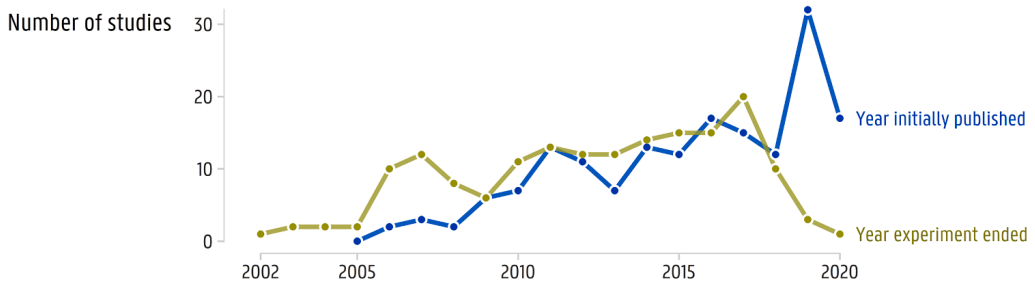


Fig. 2. Time trend of the number of studies based upon correspondence experiments. Notes. 'Year initially published' is the year in which the study was first published (as a pre-print, early-access article, or a full journal article). This year is used in our research as a criterion for study selection, while 'Year experiment ended' is used in our heterogeneity analyses as the period variable.

2019. We see two reasons for this sharp increase: (i) many correspondence experiments that ended in previous years (as early as 2013, but mostly in 2016 and 2017) were not published until 2019, and (ii) the *Journal of Ethnic and Migration Studies* compiled a special issue on ethnic discrimination in the labour market that was first published online in 2019. Logically, there is a lag between the year an experiment ends and the year the study is published. On average, this lag is 2.82 years ($SD = 2.06$). While we used the so-called 'Year initially published' as the time-related eligibility criterion in our study selection, the 'Year experiment ended' is used in further heterogeneity analyses because it constitutes a more accurate representation of the timing of a correspondence experiment (see Section 3.4.3).

In our meta-analysis, we focus on two other grouping variables besides time, namely region and call-back classification (see Section 2.4.2). Fig. 3 represents the number of correspondence experiments (i.e. units of observation) by region. The bulk of correspondence experiments is conducted in Europe ($N = 196$, 64.05%), of which 95 are in Western Europe and 60 in Northern Europe, and the Americas ($N = 75$, 24.51%), of which 64 are in Northern America (i.e. mostly the United States). Fig. 4 shows the number of correspondence audits by call-back classification. In the majority of correspondence experiments ($N = 205$, 66.99%), the authors report call-backs in the 'narrow' sense (i.e. an invitation to interview), while call-backs in the 'broad' sense (i.e. any positive response from the employer, such as a request for additional information) are reported in 101 experiments (33.01%). A detailed overview of frequencies and proportions by treatment group and region can be found in Appendix A (Table A2–A3).

Fig. 5 illustrates that the majority of experiments provide results related to the discrimination grounds of race, ethnicity, and national origin ($N = 143$, 46.73%) and gender and motherhood status ($N = 72$, 23.53%). Moreover, relying on counts, there are two discernible patterns concerning the overall treatment effect. First, for most discrimination grounds, there seems to be unequal treatment of applicants from the minority (treatment) group when compared with their majority counterparts. Second, the overall treatment of female gender applicants (vis-à-vis male gender applicants) appears highly ambiguous; in the lion's share of the experiments ($N = 33$, 53.23%), empirical evidence for unequal treatment is absent, while there is hiring discrimination against males and females in the remaining experiments. In Sections 3.2 and 3.3, we meta-analytically assess these treatment effects per discrimination ground and treatment group and address some of the heterogeneity in the uncovered hiring discrimination.

3.2. Differences in hiring discrimination by discrimination ground

Table 2 includes the pooled discrimination ratios of the correspondence experiments in our meta-analysis. Unless otherwise indicated, the findings referenced in this section (as well as Sections 3.3 and 3.4) are robust for controlling for outliers.[16,17] Detailed results of these robustness analyses can be found in Appendix A (Table A5–A6 and Table A8–A9). A list of outliers that were removed from the outlier-adjusted statistics can be retrieved from Table A12. In line with the count of votes in Section 3.1, we find empirical evidence for unequal treatment in hiring concerning the discrimination grounds of race, ethnicity, and national origin, age, religion, disability, physical appearance, wealth, and marital status. We also find some evidence of hiring discrimination regarding gender and motherhood status and sexual orientation. However, the uncovered unequal treatment based on gender and motherhood status is very small ($\widehat{DR} = 1.0413$, $CI_{95\%} = [1.0151; 1.0682]$; see also Section 3.3.2) and hiring discrimination related to sexual orientation is not robust when controlling for outliers (*k-adj.* $\widehat{DR} = 0.9007$, $CI_{95\%} = [0.7845; 1.0341]$; see Table A5 and Table A12).[18] Finally, relying on

[16] We defined outliers as those discrimination ratios for which the upper (lower) bound of the 95% confidence interval was lower (higher) than the lower (upper) bound of the confidence interval of the pooled discrimination ratio (see section 2.4.4 for more details).

[17] Different from previous versions of this paper (e.g. the IZA discussion paper; Lippens et al., 2021) and as a general rule, if $k \geq 10$, bias-adjusted estimates are reported, otherwise the original estimates are provided (Harrer et al., 2021). These bias-adjusted estimates account for potential small-study publication bias as small-sample studies might bias the original estimates because they are at greater risk of only being reported or published if they produce large, statistically significant effects compared to large-sample studies (see section 2.4.3).

[18] The abbreviation '*k*-adj.' is short for '*k*-adjusted', which indicates that a lower number of studies were included in the analysis, adjusting for influential cases (i.e. outliers). An overview of the specific outliers that were removed (by type of analysis) can be retrieved from Table A12.

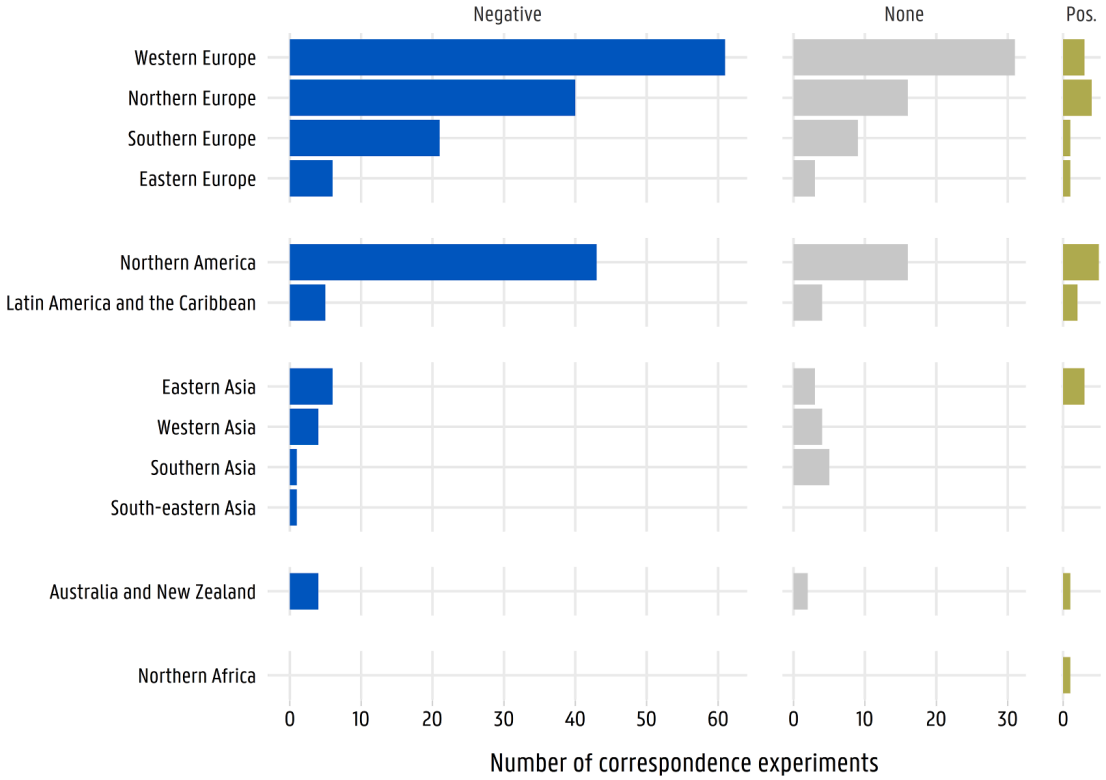


Fig. 3. Number of correspondence experiments by region (rows) and treatment effect (panels). Notes. 'Number of correspondence experiments' represents the units of observation included in the meta-analysis. The bars are grouped in panels, representing the overall treatment effect in the original correspondence experiments. Regional classification is based on the United Nations (2021) M49 Standard. Abbreviations used: Pos. (Positive).

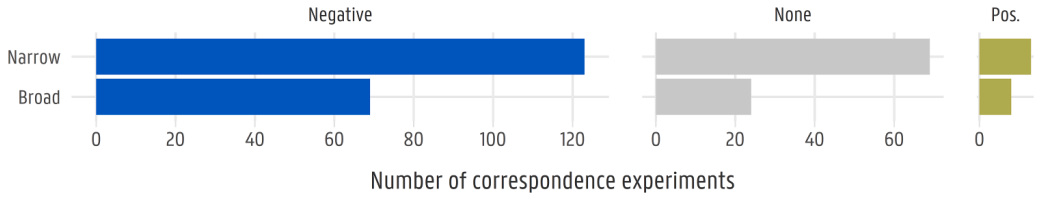


Fig. 4. Number of correspondence experiments by call-back classification (rows) and treatment effect (panels). Notes. Outcome variables consisting of an invitation to a job interview (or any broadly defined positive response from the employer, such as a request for additional information) are classified as narrow (or broad). 'Number of correspondence experiments' represents the units of observation included in the meta-analysis. Bars are grouped in panels, representing the overall treatment effect in the original correspondence experiments. Abbreviations used: Pos. (Positive).

estimates from just four experiments, we find no overall evidence of hiring discrimination based on military service or affiliation ($\widehat{DR} = 0.9983$, $CI_{95\%} = [0.7766; 1.2834]$).

The pooled discrimination ratios enable us to compare the severity of unequal treatment in hiring across different discrimination grounds. Based on these point estimates, people with disabilities are on average approximately 41% less likely to receive a positive response to a job application ($\widehat{DR} = 0.5885$, $CI_{95\%} = [0.5277; 0.6563]$), while estimates based on physical appearance and age indicate reduced positive responses by approximately 37% ($\widehat{DR} = 0.6308$, $CI_{95\%} = [0.4738; 0.8397]$) and 31% ($\widehat{DR} = 0.6867$, $CI_{95\%} = [0.6503; 0.7250]$), respectively. This contrasts with the discrimination ratios for marital status ($\widehat{DR} = 0.8846$, $CI_{95\%} = [0.8109; 0.9650]$) and wealth ($\widehat{DR} = 0.8806$, $CI_{95\%} = [0.8081; 0.9596]$), which are significantly different from but closer to one. Although we must note that for pooled discrimination ratios that rely on few studies (e.g. $k < 10$), the uncertainty around these estimates can become large and the overall effect ambiguous—we urge caution in interpreting the discrimination ratios in these cases. Most notably, in recent years, many research efforts have focused on examining hiring discrimination based on race, ethnicity, and national origin. Ethnic minority candidates face on average approximately 29% fewer positive responses ($k = 143$, 46.73% of total units of observation; $\widehat{DR} = 0.7113$,

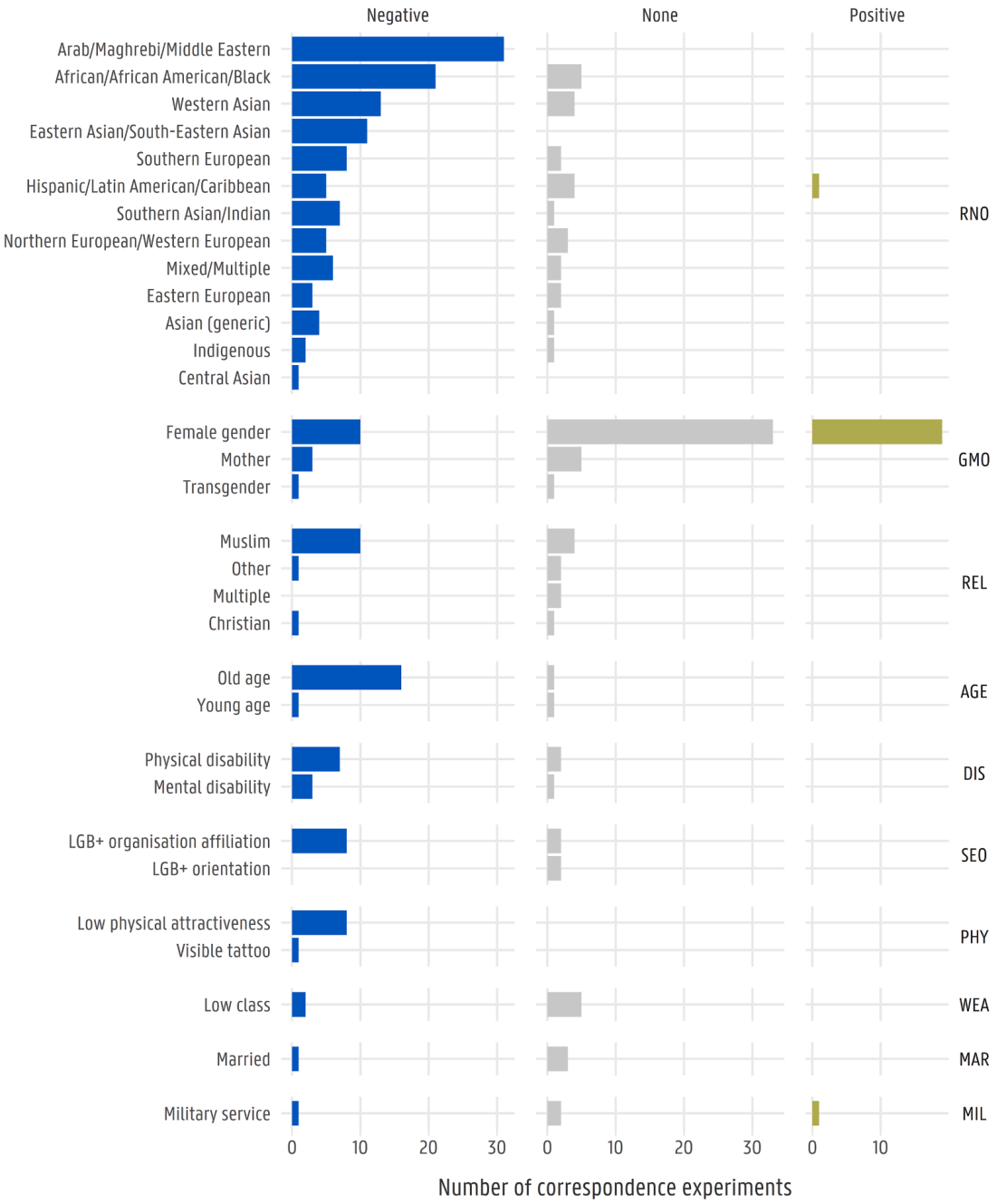


Fig. 5. Number of correspondence experiments by treatment group (rows) and treatment effect (panels). Notes. 'Number of correspondence experiments' represents the units of observation included in the meta-analysis. The bars are grouped in panels, representing the overall treatment effect in the original correspondence experiments. Abbreviations used: RNO (race, ethnicity, and national origin), GMO (gender and motherhood status), AGE (age), REL (religion), DIS (disability), SEO (sexual orientation), PHY (physical appearance), WEA (wealth), MIL (military service or affiliation), and MAR (marital status).

$CI_{95\%} = [0.6924; 0.7307]$). Nonetheless, the unequal treatment of disabled, older, and less physically attractive candidates appears at least equally problematic.[19]

In terms of statistical heterogeneity, we witness high variability in the underlying distribution of true discrimination ratios.

[19] As noted by an anonymous reviewer, it is easier to manipulate ethnicity or gender in correspondence experiments because, in many cases, the researcher merely has to change the name on the resume. In contrast, for several other discrimination grounds, the researcher would have to add organisation affiliations (e.g. for sexual orientation or military affiliation), integrate different levels of work experience (e.g. for age) or manipulate photographs (e.g. for physical appearance). This could be one of the reasons why the research focus in recent experimental research on hiring discrimination has been on ethnicity and not on other discrimination grounds.

Table 2
Pooled discrimination ratios by discrimination ground and treatment group.

Variable Discrimination ground or treatment group	Effect k	N_{obs}	N_{events}	$\widehat{DR}$ [$CI_{95\%}$]	t (p)	Statistical heterogeneity I^2
Race, ethnicity, and national origin	143	340,262	68,946	0.7113 [0.6924; 0.7307]	−24.79*** (<0.001)	90.13%
Arab/Maghrebi/Middle Eastern	31	69,311	14,523	0.5937 [0.5548; 0.6353]	−15.09*** (<0.001)	87.37%
African/African American/Black	26	69,177	12,796	0.6845 [0.6444; 0.7270]	−12.32*** (<0.001)	88.44%
Western Asian	17	34,828	7213	0.7508 [0.6977; 0.8080]	−7.66*** (<0.001)	68.13%
Eastern Asian/South-Eastern Asian	11	43,688	5528	0.6286 [0.5368; 0.7361]	−5.77*** (<0.001)	93.06%
Hispanic/Latin American/Caribbean	10	15,344	3632	0.9220 [0.8091; 1.0507]	−1.22 (0.223)	80.52%
Southern European	10	19,648	4036	0.6673 [0.5711; 0.7798]	−5.09*** (<0.001)	76.97%
Mixed/Multiple	8	14,157	3219	0.6757 [0.4287; 1.0651]	−2.04 (0.081)	86.22%
Southern Asian/Indian	8	18,987	4200	0.7004 [0.6352; 0.7723]	−8.61*** (<0.001)	53.16%
Northern European/Western European	8	12,983	3135	0.8154 [0.6661; 0.9981]	−2.39* (0.048)	76.49%
Asian (generic)	5	7098	3114	0.6739 [0.4530; 1.0024]	−2.76 (0.051)	77.77%
Eastern European	5	12,891	2907	0.7206 [0.5271; 0.9851]	−2.91* (0.044)	92.31%
Indigenous	3	20,189	3906	0.7793 [0.4127; 1.4715]	−1.69 (0.233)	95.71%
Central Asian	1	1961	737	N/A	N/A	N/A
Gender and motherhood status	72	330,600	56,650	1.0413 [1.0151; 1.0682]	3.11** (0.002)	94.07%
Female gender	62	308,840	53,309	1.0413 [1.0138; 1.0696]	2.97** (0.003)	94.81%
Mother	8	19,394	2471	0.9044 [0.7887; 1.0370]	−1.74 (0.126)	30.49%
Transgender	2	2366	870	0.8500 [0.5306; 1.3619]	−4.38 (0.143)	N/A
Age	19	86,730	11,775	0.6867 [0.6503; 0.7250]	−13.54*** (<0.001)	82.36%
Old age	17	82,642	11,220	0.6646 [0.6292; 0.7020]	−14.64*** (<0.001)	83.53%
Young age	2	4088	555	0.7698 [0.2294; 2.5830]	−2.75 (0.222)	3.65%
Religion	21	41,917	11,447	0.7855 [0.7457; 0.8274]	−9.11*** (<0.001)	92.45%
Muslim	14	24,344	4947	0.7730 [0.7069; 0.8452]	−5.65*** (<0.001)	85.73%
Other	3	9652	3684	0.8240 [0.3578; 1.8979]	−1.00 (0.423)	91.32%
Christian	2	4642	1375	0.7293 [0.0075; 71.1483]	−0.88 (0.542)	97.75%
Multiple	2	3279	1441	0.9275 [0.7532; 1.1422]	−4.59 (0.137)	N/A
Disability	13	25,232	4530	0.5885 [0.5277; 0.6563]	−9.53*** (<0.001)	96.80%
Physical disability	9	19,694	4191	0.5369 [0.2607; 1.1056]	−1.99 (0.082)	97.83%
Mental disability	4	5538	339	0.6249 [0.4075; 0.9581]	−3.50* (0.039)	38.43%
Sexual orientation[a]	12	41,763	13,442	0.7016 [0.5138; 0.9581]	−2.50* (0.029)	98.53%
LGB+ organisation affiliation[a]	10	38,520	13,018	0.6482 [0.4539; 0.9257]	−2.75* (0.022)	98.78%
LGB+ orientation	2	3243	424	1.0585 [0.5470; 2.0485]	1.09 (0.471)	N/A
Physical appearance	9	50,070	9981	0.6308 [0.4738; 0.8397]	−3.71** (0.006)	97.88%
Wealth	7	11,517	1903	0.8806 [0.8081; 0.9596]	−3.62* (0.011)	N/A
Marital status	4	18,369	2715	0.8846 [0.8109; 0.9650]	−4.49* (0.021)	N/A
Military service or affiliation	4	18,208	1738	0.9983 [0.7766; 1.2834]	−0.02 (0.985)	67.43%

Notes. Abbreviations and notations used: k (number of correspondence experiments), N_{obs} (total number of observations), N_{events} (total number of positive call-backs), $\widehat{DR}$ (pooled discrimination ratio estimate), $CI_{95\%}$ (95% confidence interval), $\widehat{DR}^{*}$ $CI_{95\%}$, LGB+ (lesbian, gay, and bisexual, amongst other sexual orientations), and N/A (not applicable). Pooled discrimination rates are only calculated for discrimination grounds or treatment groups for which $k > 1$. If $k \geq 10$, bias-adjusted estimates from the 'limit' meta-analyses are reported, otherwise the original estimates are provided (see Section 2.4.3). Following Schwarzer et al. (2015), statistical heterogeneity statistics are calculated for those grounds or groups for which $k > 2$. Following Higgins and Thompson (2002), I^2 values around 25%, 50%, or 75% indicate low, moderate, or high heterogeneity, respectively. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

[a] Because the statistical heterogeneity related to this pooled discrimination ratio is extremely high (i.e. I^2 equals approximately 1), resulting in an inaccurate bias-adjusted estimate, we report the unadjusted estimate instead of the publication bias-adjusted estimate obtained from the 'limit' analysis. Here, we thus make an exception to the rule that for analyses where $k \geq 10$ the adjusted estimate is reported

Specifically, I^2 estimates range from 82.36% (for age) to 98.53% (for sexual orientation)—not considering the exceptional cases of wealth, military service or affiliation, and marital status, which are based on a too low number of correspondence experiments to meaningfully interpret the statistical heterogeneity. This signals that the findings of the experiments clustered within the respective discrimination grounds are highly disparate.[20] However, this is not surprising: similar estimates are expected when pooling the discrimination ratios in such broad categories. In Sections 3.3 and 3.4, we assess design-related heterogeneity based on treatment group, call-back classification, region, and period, which helps pinpoint whether this large underlying statistical variability can be (partly) explained by discrepancies based on these variables across study designs.

3.3. Differences in hiring discrimination by treatment group

In this section, we examine differences in hiring discrimination by treatment group. This analysis provides a more granular view of

[20] This confirms our priors that (i) there is between-study variation that is driven by differences in the operationalisation of the experimental groups and, more broadly, how the correspondence experiments are designed, and (ii) the random-effects model is an appropriate model to account for such between-study heterogeneity.

the pooled discrimination ratios described above, as the pooling of said ratios at the level of the discrimination ground substantially masks relevant information about their underlying variability at the level of the treatment group. Estimates by treatment group are also given in Table 2. Moreover, Fig. 6 illustrates visually the relative change in the probability of a positive call-back for the applicants belonging to the respective treatment groups vis-à-vis their counterparts in the control group.

3.3.1. Race, ethnicity, and national origin

First of all, regarding race, ethnicity, and national origin, unfavourable treatment in hiring is highest for applicants belonging to the groups of Arab, Maghrebi, or Middle Eastern ($\widehat{DR} = 0.5937$, $CI_{95\%} = [0.5548; 0.6353]$), Eastern Asian or South-Eastern Asian ($\widehat{DR} = 0.6286$, $CI_{95\%} = [0.5368; 0.7361]$), and Southern European ($\widehat{DR} = 0.6673$, $CI_{95\%} = [0.5711; 0.7798]$) applicants. On average, these applicants face approximately a 41%, 37%, and 33% reduction in the probability of a positive call-back, respectively. We also find overall hiring discrimination against African, African American, or Black ($\widehat{DR} = 0.6845$, $CI_{95\%} = [0.6444; 0.7270]$), Southern Asian or Indian ($\widehat{DR} = 0.7004$, $CI_{95\%} = [0.6352; 0.7723]$, and Western Asian ($\widehat{DR} = 0.7508$, $CI_{95\%} = [0.6977; 0.8080]$) applicants. Furthermore, of all European treatment groups, Southern Europeans experience hiring discrimination to the largest extent, while the discrimination ratios related to the applicants of Eastern European origin ($\widehat{DR} = 0.7206$, $CI_{95\%} = [0.5271; 0.9851]$) or (White) Northern and Western origin ($\widehat{DR} = 0.8154$, $CI_{95\%} = [0.6661; 0.9981]$) are closer to one, meaning that they face less discrimination.[21,22]

Perhaps more surprisingly, at first, we do not find evidence for overall hiring discrimination against Hispanic, Latin American, or Caribbean applicants ($\widehat{DR} = 0.9220$, $CI_{95\%} = [0.8091; 1.0507]$), despite several individual correspondence experiments providing evidence for the unequal treatment of these applicants in hiring (see Fig. 5). However, when the only identified outlier from the analysis of this minority group is excluded, the pooled discrimination ratio becomes statistically significant at the 5% level (*k-adj.* $\widehat{DR} = 0.8175$, $CI_{95\%} = [0.7095; 0.9420]$; see Table A6 and Table A12). This result is in line with previous review studies by Quillian et al. (2017, 2019), who found that discrimination against applicants of Latin American origin seemed to be generally lower than discrimination against applicants belonging to Black, Middle Eastern, North African, or Asian minority groups.

3.3.2. Gender and motherhood status

Next, we take a closer look at hiring discrimination based on gender and motherhood status. We observe a slightly positive discrimination ratio regarding the female gender; there is an average 4.13% higher probability for female gender candidates of receiving a positive response to an application ($\widehat{DR} = 1.0413$, $CI_{95\%} = [1.0138; 1.0696]$). In contrast, we find no evidence of hiring discrimination related to the treatment groups 'transgender' or 'mother'. After excluding outliers from the analysis, the discrimination ratio for female gender (vis-à-vis male gender) applicants remains statistically significant and positive (*k-adj.* $\widehat{DR} = 1.0663$, $CI_{95\%} = [1.0221; 1.1124]$; see Table A6 and Table A12). However, we note that the statistical heterogeneity related to this pooled discrimination ratio is very high ($I^2 = 94.81\%$; see Table 2)—this is exemplified visually in Fig. 6. As pointed out in Section 3.1, we already know that the majority of correspondence experiments on gender discrimination in hiring find null results while the remaining studies find hiring discrimination against male gender candidates but also against female gender candidates (see Fig. 5).

The above findings are in line with a recent large-scale correspondence audit study conducted in the United States (with more than 80,000 applications) in which the authors found that contact rates for male gender and female gender applicants differed significantly between firms: some employers favoured male gender candidates, while others favoured female gender candidates (Kline et al., 2021). Similar findings arise from another recent large-scale correspondence audit study in Australia (with more than 12,000 applications) where males received substantially more (less) positive responses than females in occupations dominated by male (female) gender workers (Adamovic and Leibbrandt, 2022). In our analysis, we do not go further into what could drive this variability. One explanation is that demand-side factors, such as the influence of certain male- or female-orientated job characteristics on the selection criteria used by employers, may lead to gender-based hiring, but also the prototypicality of a candidate for their gender category (Cortina et al., 2021; Yavorsky, 2019). Van Borm and Baert (2022), for example, found women to be perceived as more social and supportive but less assertive or physically strong than men. Applications from female gender candidates for caregiving job positions could consequently be positively received by employers or recruiters, while the opposite might be true for male gender applicants. Another explanation is that

[21] Applicants with Albanian-sounding names are discriminated against in Greece and Italy, applicants with Greek names are unfavourably treated in Canada, applicants of Italian origin are discriminated against in Australia and Belgium, and applicants with a Serbian name and appearance experience hiring discrimination in Austria. The control group always consisted of same-country applicants belonging to their region's majority ethnic group.

[22] The Northern and Western European treatment groups comprised minority (majority) applicants of English (Finnish), French (German), German (Irish, Italian, or Russian), and Latvian or Lithuanian (Russian) origin for whom unequal treatment in hiring was assessed. The Eastern European treatment group consisted of minority (majority) applicants of Russian (Finnish), Romanian (Italian), Ukrainian (Russian or Greek), and Polish (Swedish) origin.

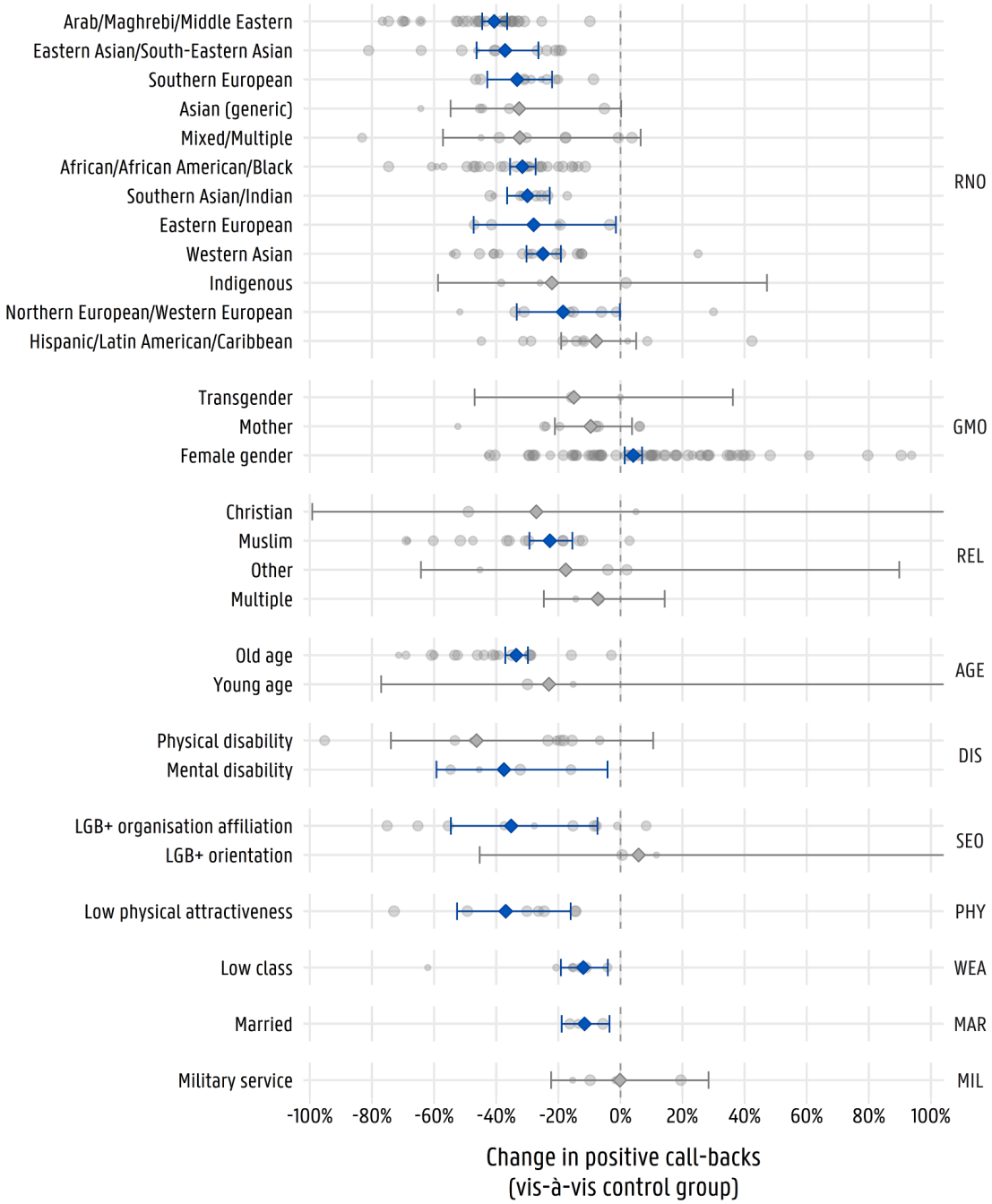


Fig. 6. Hiring discrimination based on pooled discrimination ratios by treatment group. Notes. The change in positive call-backs (compared to the control group), represented by filled diamond shapes, is calculated by subtracting one from the corresponding pooled discrimination ratio (i.e. $\widehat{DR}$; see Table 2). Error bars illustrate the 95% confidence intervals of these ratios. Statistically insignificant ratios (at the 5% level) are greyed out. Semi-transparent dots represent the discrimination ratios of the individual correspondence experiments. Abbreviations used: RNO (race, ethnicity, and national origin), GMO (gender and motherhood status), AGE (age), REL (religion), DIS (disability), SEO (sexual orientation), PHY (physical appearance), WEA (wealth), MAR (marital status), and MIL (military service or affiliation). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the self-selection of members of one gender group into specific sectors or jobs, creating a predominance of that gender in those sectors or jobs, could lie at the root of discrimination against members of the opposite group (Cortina et al., 2021).

3.3.3. Religion

The majority of correspondence experiments on religion discrimination in hiring have focused on Muslims ($k = 14$, 66.67% of the total).[23] The remaining correspondence experiments have looked at a highly diverse subset of religions (i.e. evangelical, Jehovah's Witness, Pentecostal, Christian [generic], Buddhist, Hindu, Jewish, no religious affiliation, and various religious affiliations simultaneously). Hiring discrimination based on religion seems to be mainly driven by the unequal treatment of this former group (i.e. Muslims; $\widehat{DR} = 0.7730$, $CI_{95\%} = [0.7069; 0.8452]$). Because of the low density of correspondence experiments concerning other religions, the estimates regarding these treatment groups bear limited reliability. The latter is exemplified by the broad confidence intervals around these estimates (see Table 2 and Fig. 6).

3.3.4. Age

Older applicants (vis-à-vis young to middle-aged applicants), but not younger applicants (vis-à-vis middle-aged applicants), are strongly discriminated against in correspondence experiments ($\widehat{DR} = 0.6646$, $CI_{95\%} = [0.6292; 0.7020]$). Yet, the estimate for younger applicants is based on a very small number of tests ($k = 2$), which warrants caution in interpreting this discrimination ratio. Important to note is that the operationalisation of age differed substantially across correspondence experiments. In this respect, we refer the reader to Appendices A (Table A4) and B (Table B1) for details concerning what constituted 'older' and 'younger' applicants in the original studies. A representative example is the study by Riach (2015), where older candidates were 47 years old, while younger candidates were 27 years old, creating a 20-year age gap between the treatment group and the control group. However, some experiments only look at a 6-year age gap (e.g. Baert et al., 2016), while others go as far as examining a 35-year age gap (e.g. Neumark et al., 2019).

3.3.5. Disability

Discrimination based on disability seems to be equally prompted by the unequal treatment of applicants with a physical disability ($\widehat{DR} = 0.5369$, $CI_{95\%} = [0.2607; 1.1056]$) or a mental disability ($\widehat{DR} = 0.6249$, $CI_{95\%} = [0.4075; 0.9581]$). Although the original estimate related to physical disability is not statistically significant, this discrimination ratio becomes statistically significant at the 5% level after excluding one outlier from the analysis (*k-adj.* $\widehat{DR} = 0.7494$, $CI_{95\%} = [0.6259; 0.8974]$; see Table A6 and Table A12). Here, too, there is a large variety of what is considered a physical disability or mental disability in the original studies (see Table A4 and Table B1). Physical disability includes obesity, blindness or deafness, HIV infection, spinal cord injury, being a wheelchair user, or an unspecified physical disability. Mental disability comprises Asperger's Syndrome, autism, former depression, or a history of mental illness. Given the broad operationalisation of physical disability, it is not surprising that the statistical heterogeneity linked to the pooled discrimination ratio is consequently very high ($I^2 = 97.83\%$; see Table 2 and Fig. 6).

3.3.6. Sexual orientation

Finally, the results concerning hiring discrimination based on sexual orientation are somewhat mixed. The main effect is primarily driven by correspondence experiments considering individuals who have an affiliation with an LGB+ organisation (e.g. membership in an LGB+ rights organisation; $\widehat{DR} = 0.6482$, $CI_{95\%} = [0.4539; 0.9257]$, $k = 10$) in comparison with those who directly disclose an LGB+ orientation ($\widehat{DR} = 1.0585$, $CI_{95\%} = [0.547; 2.0485]$, $k = 2$). Although this finding could raise the question of whether hiring discrimination based on sexual orientation is mainly motivated by a discriminatory stance against activism (i.e. affiliation with an organisation that supports LGB+ rights) rather than discriminatory attitudes regarding LGB+ orientation per se (see also Baert, 2014), the basis for comparison is very narrow. More specifically, we are comparing the results of ten experiments where LGB+ orientation is signalled through an affiliation with those of only two experiments where the LGB+ orientation is disclosed directly.[24] After excluding two clear outliers from the analysis, we also observe that the unequal treatment based on affiliation with an LGB+ organisation becomes statistically insignificant at the 5% level (*k-adj.* $\widehat{DR} = 0.7924$, $CI_{95\%} = [0.6203; 1.0122]$; see Table A6 and Table A12). Overall, the pooled estimates related to sexual orientation seem to be impacted by some form of publication bias (see Table A11 and Figure A2–16).

[23] As pointed out by Bartkoski et al. (2018) and Di Stasio et al. (2021), it is important to make the distinction between hiring discrimination that is purely due to religion (e.g. against Muslims) and hiring discrimination that is due to a combination of origin and religion (e.g. against Arabs). This distinction may have been overseen in previous correspondence experiments. Potentially, in correspondence experiments where the effects of origin and religion interact, confounding might lead to a wrong estimation of the true discrimination. Di Stasio et al. (2021) have attempted to disentangle these effects in their experiment by considering discrimination against 'disclosed Muslims' (i.e. a religion effect) separately from discrimination against 'Muslims by default' (i.e. a religion and/or origin effect).

[24] As an anonymous reviewer has pointed out, other studies focusing on different discrimination grounds could face the same criticism—i.e. the signal used may not be externally valid. For example, in the study of Ameri et al. (2018), disability is signalled in a similar way. Considering race, ethnicity, and national origin, there is an active discussion that African-American or Hispanic names, used to signal African-American or Hispanic status, may also signal socio-economic status (e.g. Darolia et al., 2016; Gaddis, 2017). Moreover, in many of the studies where LGB+ orientation status is signalled through activism or affiliation, there is a similar non-LGB+ signal that is attributed to the control group, which might 'wash out' the activism or affiliation effect (e.g. Drydakis, 2009; Tilcsik, 2011).

Table 3
Differences in pooled discrimination ratios by call-back classification, region, and period per discrimination ground.

Variable			Effect					Statistical heterogeneity
Discrimination ground	Sub-group	Level	k	N_{obs}	N_{events}	$\widehat{DR}$ [$CI_{95\%}$]	t (p)	I^2
Race, ethnicity, and national origin	Call-back	Narrow	87	169,762	35,019	0.6953 [0.6694; 0.7222]	−18.76*** (<0.001)	86.97%
		Broad	56	170,500	33,927	0.7865 [0.7561; 0.8181]	−11.95*** (<0.001)	92.43%
	Region	Americas	38	114,102	17,276	0.8027 [0.7618; 0.8457]	−8.24*** (<0.001)	88.60%
		Europe	94	175,525	46,261	0.7004 [0.6781; 0.7234]	−21.56*** (<0.001)	89.34%
		Asia	6	40,796	3130	0.4877 [0.2900; 0.8203]	−3.55* (0.016)	93.03%
		Other	5	9839	2279	0.6275 [0.4115; 0.9570]	−3.07* (0.037)	90.95%
	Period	2002–2010	45	80,977	18,016	0.6939 [0.6558; 0.7344]	−12.65*** (<0.001)	78.42%
		2011–2020	98	259,285	50,930	0.7211 [0.6992; 0.7436]	−20.81*** (<0.001)	91.57%
Gender and motherhood status	Call-back	Narrow	52	192,641	36,247	1.0766 [1.0449; 1.1092]	4.84*** (<0.001)	87.70%
		Broad	20	137,959	20,403	0.9335 [0.8877; 0.9817]	−2.68** (0.007)	96.87%
	Region	Americas	10	93,335	13,782	1.0498 [0.9971; 1.1053]	1.85 (0.064)	98.20%
		Europe	48	153,130	29,199	1.0190 [0.9840; 1.0552]	1.06 (0.291)	89.92%
		Asia	11	76,260	11,495	1.0422 [0.9795; 1.1089]	1.31 (0.191)	80.37%
		Other	3	7875	2174	1.2854 [0.8355; 1.9775]	2.51 (0.129)	46.28%
	Period	2002–2010	23	93,799	14,824	1.1440 [1.0647; 1.2293]	3.67*** (<0.001)	82.53%
		2011–2020	49	236,801	41,826	0.9878 [0.9612; 1.0151]	−0.88 (0.377)	95.07%
Age	Call-back	Narrow	14	34,541	3549	0.7160 [0.6118; 0.8380]	−4.16*** (<0.001)	82.52%
		Broad	5	52,189	8226	0.6248 [0.5038; 0.7749]	−6.07** (0.004)	74.04%
	Region	Americas	6	61,411	8581	0.6881 [0.6438; 0.7354]	−14.44*** (<0.001)	29.92%
		Europe	13	25,319	3194	0.6288 [0.5349; 0.7392]	−5.62*** (<0.001)	78.87%
		Asia	N/A	N/A	N/A	N/A	N/A	N/A
		Other	N/A	N/A	N/A	N/A	N/A	N/A
	Period	2002–2010	6	19,980	1388	0.5460 [0.4149; 0.7185]	−5.67** (0.002)	52.05%
		2011–2020	13	66,750	10,387	0.6828 [0.6439; 0.7240]	−12.75*** (<0.001)	86.45%
Religion	Call-back	Narrow	11	19,002	4158	0.8350 [0.7484; 0.9317]	−3.23** (0.001)	91.71%
		Broad	10	22,915	7289	0.7425 [0.6984; 0.7894]	−9.53*** (<0.001)	92.91%
	Region	Americas	5	9494	1409	0.7535 [0.5108; 1.1117]	−2.02 (0.113)	81.97%
		Europe	14	26,957	9785	0.8045 [0.7588; 0.8530]	−7.29*** (<0.001)	94.08%
		Asia	2	5466	253	0.6336 [0.0015; 265.6794]	−0.96 (0.513)	92.74%
		Other	N/A	N/A	N/A	N/A	N/A	N/A
	Period	2002–2010	4	7273	1489	0.5633 [0.2738; 1.1588]	−2.53 (0.085)	77.96%
		2011–2020	17	34,644	9958	0.8199 [0.7756; 0.8668]	−7.00*** (<0.001)	90.16%

(*continued on next page*)

Table 3 (*continued*)

Variable			Effect					Statistical heterogeneity
Discrimination ground	Sub-group	Level	k	N_{obs}	N_{events}	$\widehat{DR}$ [$CI_{95\%}$]	t (p)	I^2
Disability	Call-back	Narrow	11	22,645	3580	0.5332 [0.4643; 0.6123]	−8.91*** (<0.001)	97.30%
		Broad	2	2587	950	0.7586 [0.4853; 1.1859]	−7.86 (0.081)	N/A
	Region	Americas	5	13,382	1198	0.7308 [0.5198; 1.0275]	−2.56 (0.063)	70.41%
		Europe	8	11,850	3332	0.4722 [0.2110; 1.0569]	−2.20 (0.064)	98.02%
		Asia	N/A	N/A	N/A	N/A	N/A	N/A
		Other	N/A	N/A	N/A	N/A	N/A	N/A
	Period	2002–2010	3	7494	2035	0.3159 [0.0054; 18.5733]	−1.22 (0.348)	99.40%
		2011–2020	10	17,738	2495	0.8017 [0.7080; 0.9077]	−3.49*** (<0.001)	61.76%
Sexual orientation[a]	Call-back	Narrow	10	35,807	12,407	0.6534 [0.4540; 0.9404]	−2.64* (0.027)	98.77%
		Broad	2	5956	1035	0.9970 [0.3606; 2.7565]	−0.04 (0.976)	51.61%
	Region	Americas	3	7179	777	0.7650 [0.4181; 1.4000]	−1.91 (0.197)	75.81%
		Europe	8	30,058	11,220	0.7735 [0.5378; 1.1126]	−1.67 (0.139)	97.78%
		Asia	1	4526	1445	0.2489 [0.2218; 0.2793]	−23.67*** (<0.001)	N/A
		Other	N/A	N/A	N/A	N/A	N/A	N/A
	Period	2002–2010	5	17,678	3719	0.6159 [0.3468; 1.0936]	−2.34 (0.079)	97.84%
		2011–2020	7	24,085	9723	0.7718 [0.4758; 1.2520]	−1.31 (0.238)	98.73%

Notes. Abbreviations and notations used: k (number of correspondence experiments), N_{obs} (total number of observations), N_{events} (total number of positive call-backs), $\widehat{DR}$ (pooled discrimination ratio estimate), $CI_{95\%}$ (95% confidence interval), $\widehat{DR}^{*}$ $CI_{95\%}$, and N/A (not applicable). 'Narrow' refers to correspondence experiments in which the call-back variable is related to an invitation to a job interview; 'broad' refers to experiments in which said variable conveys any positive reaction to an application (e.g. an employer's request for additional. Discrimination rates at the sub-group level are only calculated for discrimination grounds for which the combined number of correspondence experiments equals or exceeds 10. If $k \geq 10$ for a given sub-group level, bias-adjusted estimates from the 'limit' meta-analyses are reported, otherwise the original estimates are provided (see Section 2.4.3). Following Schwarzer et al. (2015), statistical heterogeneity statistics are only calculated for those grounds for which $k > 2$. Following Higgins and Thompson (2002), I^2 values around 25%, 50%, or 75% indicate low, moderate, or high heterogeneity, respectively. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

[a] Because the statistical heterogeneity related to these pooled discrimination ratios is extremely high (i.e. I^2 equals approximately 1 for some estimates), resulting in an inaccurate bias-adjusted estimate, we report the unadjusted estimate instead of the publication bias-adjusted estimate obtained from the limit analysis. Here, we thus make an exception to the rule that for analyses where $k \geq 10$ the adjusted estimate is reported

3.4. Differences in hiring discrimination by call-back classification, region, and period

In this section, we report on the heterogeneity of our results by call-back classification, region, and period. Table 3 contains the pooled discrimination ratios by sub-group and sub-group level per discrimination ground. These ratios are discussed alongside those of the heterogeneity analyses by treatment group, which are included in Appendix A (Table A7). Moreover, Table 4 contains the results from the weighted least squares meta-regression of hiring discrimination on call-back classification, region, and period per discrimination ground. Analogously, the results from similar meta-regression analyses at the level of the treatment group can also be retrieved from Appendix A (Table A14–A15).

3.4.1. Call-back classification heterogeneity

In the hiring discrimination literature, the classification of the outcome measure (i.e. call-back) is usually approached in two ways. Either the reported call-back is measured in the broad sense, where any positive response to the application is interpreted as a positive response, or the call-back is measured in the narrow sense, where only an invitation to interview is regarded as a positive response (but sometimes also both). We assume that hiring discrimination may be lower if the call-back is measured in the broad sense compared to the narrow sense. On the one hand, when considering call-back in the broad sense, this measure also includes so-called 'courtesy questions' where the recruiter or employer poses a question to the minority candidate (but not to the majority candidate) without

ring discrimination on call-back classification, region, and period.

	GMO	AGE	REL	DIS	SEO
.1630* (15.4015)	19.9726 (12.9508)	−19.0039 (25.6796)	−28.7969 (78.6753)	−59.1381 (136.8726)	−180.4927 (93.3485)
	N/A	*N/A*	*N/A*	*N/A*	*N/A*
594 (0.0637)	0.0123 (0.0820)	−0.1189 (0.0944)	−0.0370 (0.2555)	0.2147 (0.2538)	0.3013 (0.2043)
	N/A	*N/A*	*N/A*	*N/A*	*N/A*
	0.7105*** (0.1268)	N/A	N/A	N/A	N/A
906 (0.2133)	0.0743 (0.1293)	N/A	−0.1604 (0.6872)	N/A	−1.1651* (0.2440)
833 (0.0687)	0.0356 (0.1192)	−0.2839* (0.0974)	−0.1127 (0.2134)	−0.2758 (0.2193)	−0.0775 (0.2763)
554 (0.3084)	0.1757 (0.1230)	N/A	N/A	N/A	N/A
73* (0.0077)	−0.0099 (0.0065)	0.0093 (0.0128)	0.0142 (0.0391)	0.0292 (0.0679)	0.0896 (0.0464)
	72	19	21	13	12
7	0.047	0.028	0.081	0.489	0.044
5%	92.11%	77.69%	93.07%	96.23%	89.41%
42	18.702	9.587	25.831	38.128	10.737
7%	7.93%	42.58%	7.80%	10.98%	79.53%

NO (race, ethnicity, and national origin), GMO (gender and motherhood status), AGE (age), REL (religion), DIS (disability), SEO (sexual orientation), not available), *k* (number of correspondence experiments or effects), *AIC* (Akaike information criterion). Following Schwarzer et al. (2015), meta-regression ion grounds for which $k \geq 10$. Presented statistics are coefficient estimates with standard errors between parentheses. Standard errors were clustered at the ignify more (less) hiring discrimination against minority candidates for a given variable or dummy category. Following Higgins and Thompson (2002), I^2 w, moderate, or high heterogeneity, respectively. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

necessarily having the intention of wanting to know the answer to that question (e.g. 'Could you elaborate a bit on your work experience?'). In these cases, it merely serves as an excuse to dismiss the minority candidate, which might, in turn, obfuscate actual discrimination. On the other hand, this assumption aligns with the idea that—based on statistical discrimination theory—the suitability of minority candidates is literally more likely to be questioned (Lippens et al., 2022). Asking questions to the candidate directly is an evident and cost-effective way to (in)validate a recruiter's or employer's possibly stereotypical image of this candidate.

Nevertheless, we find no evidence for differences in hiring discrimination by call-back classification when controlling for region and period effects (see Table 4 and Table A14). The broadness in measuring or reporting call-backs does not seem to relate to the ratio between the probability of a positive call-back for minority applicants and the probability of a positive response for majority-group candidates (i.e. the discrimination ratio). In other words, levels of hiring discrimination do not appear significantly different if the authors measure and record call-backs in the narrow sense or the broad sense.

3.4.2. Region heterogeneity

There are numerous reasons why hiring discrimination may vary across regions. Differences in terms of legislation, public policies, and socio-economic contexts, amongst other things, can lead to differences in the treatment of minority candidates. A first example, associated with race, ethnicity, and national origin, is that European countries have generally known a much larger influx of migrants from North and Sub-Saharan Africa than North America, while the opposite is true for migrants from Central America and Southern, Eastern, and South-Eastern Asia (Abel and Sander, 2014). Depending on the prevailing theoretical frame of reference, this can be (dis) advantageous to the migrated group: a large migration flow of a specific minority group can help in updating stereotypical ideas of individuals from this group (i.e. statistical discrimination) but it can also elicit prejudice against members of this out-group (i.e. taste-based discrimination; Lippens et al., 2022). A second example, related to age, is that the legal framework surrounding pensions and retirement is much stricter in European countries than it is in the United States, which may have a strong effect on social norms regarding the employment of older candidates (Lahey, 2010). Specifically, mandatory retirement ages—effective in many European countries—may signal to employers that it is appropriate for older people not to be active in the labour market anymore when they approach a certain age, inducing age discrimination in hiring.

We find several regional differences in hiring discrimination. The unfavourable treatment of ethnic minority candidates appears to be greater in Asia ($\widehat{DR} = 0.4877$, $CI_{95\%} = [0.2900; 0.8203]$) than in the Americas ($\widehat{DR} = 0.8027$, $CI_{95\%} = [0.7618; 0.8457]$) or Europe ($\widehat{DR} = 0.7004$, $CI_{95\%} = [0.6781; 0.7234]$; see Table 3). However, controlling for call-back classification and period effects, this difference is not statistically significant (see Table 4 and Table A13). Zooming in on the lower-level treatment groups, we do observe that applicants of Western Asian origin (e.g. Azeri, Armenians, Kurds, Uyghurs) experience significantly more discrimination in Asia ($\widehat{DR} = 0.5321$, $CI_{95\%} = [0.4250; 0.6661]$) than Europe ($\widehat{DR} = 0.8008$, $CI_{95\%} = [0.7427; 0.8634]$; $\widehat{\beta} = -0.3561$, $p = 0.025$; Table A7 and Table A15).[25] The former region comprises both Eastern and Western Asian countries (i.e. China, Georgia, and Turkey). These higher levels of hiring discrimination could be explained by (i) the relatively large local presence of these minority groups in Asian labour markets compared to European labour markets and (ii) the negative connotations associated with these groups within these particular regions, which do not necessarily exist in European countries (Asali et al., 2018; Maurer-Fazio, 2013).

Although absolute levels of hiring discrimination based on race, ethnicity, and national origin appear to be higher in Europe ($\widehat{DR} = 0.7004$, $CI_{95\%} = [0.6781; 0.7234]$) than in the Americas ($\widehat{DR} = 0.8027$, $CI_{95\%} = [0.7618; 0.8457]$), we find no statistically significant evidence for such heterogeneity at the (sub-)regional level controlling for call-back classification and period effects ($\widehat{\beta} = -0.0833$; $p = 0.167$; see Table 4). This contrasts with the findings of Quillian et al. (2019) and Zschirnt and Ruedin (2016) who did find higher levels of ethnic hiring discrimination in Europe vis-à-vis the Americas. The result does not change when we adjust the discrimination ratios for outliers—the gap between the estimates even narrows (see Table A8). This discrepancy is presumably because we primarily focused on differences at the (sub-)regional level and not on differences at the country level. When we compare hiring discrimination in several European countries with the United States directly, we observe that the unequal treatment of ethnic minority candidates in hiring is higher in Finland ($k = 3$, $\widehat{\beta} = -0.4890$, $p = <0.001$), France ($k = 7$, $\widehat{\beta} = -0.2720$, $p = 0.043$), and Italy ($k = 6$, $\widehat{\beta} = -0.2754$, $p = 0.023$) but lower in Germany ($k = 10$, $\widehat{\beta} = 0.2378$, $p = 0.033$) after controlling for call-back classification, period, and treatment group effects (see Table A16 and Figure A3-1 to Figure A3-2). This coincides with some of the findings of Quillian et al. (2019). Note that several of these results are based on a small number of correspondence experiments and thus should be interpreted with caution.

Moreover, hiring discrimination based on gender and motherhood status appears lower in Africa than in the Americas ($\widehat{\beta} = 0.7105$, $p = <0.001$; see Table 3). Nevertheless, this result is not generalisable because the coefficient estimate relies on the comparison of just one African correspondence experiment with ten American experiments.

We also observe regional differences in unequal treatment in hiring based on age, controlling for region and period effects. Generally, age discrimination in hiring is more severe in Europe ($\widehat{DR} = 0.6288$, $CI_{95\%} = [0.5349; 0.7392]$) than in the Americas ($\widehat{DR} = 0.6881$, $CI_{95\%} = [0.6438; 0.7354]$; $\widehat{\beta} = -0.2839$, $p = 0.034$; see Table 3 and Table 4).[26] Specifically, older applicants are more severely

[25] We did not identify any correspondence audits originating from the Americas in which applicants of Western Asian origin were considered as the treatment group.

[26] The meta-regression model concerning age accounts for a substantial amount of heterogeneity (Pseudo-R^2 = 42.58% considering the discrimination ground 'age' and Pseudo-R^2 = 57.00% considering the treatment group 'old age'; see Table 4 and Table A14). In other words, the region seems to be an important variable in explaining the variability between correspondence experiments on age discrimination in hiring.

discriminated against in various European countries (i.e. Belgium, France, the United Kingdom, Spain, and Sweden; $\widehat{DR} = 0.5152$, $CI_{95\%} = [0.4258; 0.6234]$) than in the United States ($\widehat{DR} = 0.6916$, $CI_{95\%} = [0.6342; 0.7541]$; $\widehat{\beta} = -0.3423$, $p = 0.010$; see Table A7, Table A14, and Figure A3–3 to Figure A3-4). This finding is exceptional given that the ages in the treatment groups of the European correspondence experiments range from 37 to 56 years, while the ages used in the American studies are generally higher, ranging from 50 to 66 years.[27] Nonetheless, this regional difference is in line with the average employment rate of 55- to 64-year-olds for the period 2002 (the year of the first correspondence experiments regarding age included in this review) to 2017 (the year of the last correspondence experiments regarding age included in this review) in the United States (60.97%) compared with Belgium (36.89%), France (42.32%), and the United Kingdom (59.89%; OECD, 2021). This finding is also in line with the analysis of Lahey (2010) in that the legislation and social norms around working at an advanced age are on average more lenient in the United States compared to European countries.

Last, we initially observe that hiring discrimination against applicants who are affiliated with an LGB+ organisation or who signal to be LGB+ orientated is higher in Asia ($\widehat{DR} = 0.2489$, $CI_{95\%} = [0.2218; 0.2793]$) than in the Americas ($\widehat{DR} = 0.7650$, $CI_{95\%} = [0.4181; 1.400]$; $\widehat{\beta} = -1.1651$, $p = 0.019$) or Europe ($\widehat{DR} = 0.7735$, $CI_{95\%} = [0.5378; 1.1126]$; $\widehat{\beta} = -1.0876$, $p = <0.001$; see Table 3, Table 4, and Table A13). Here, too, we controlled for call-back classification and period effects. However, we have reasons to believe that this is a spurious correlation: the discrimination ratio of the Asian region is based on just one study from Cyprus (Drydakis, 2014), which is part of Western Asia according to the United Nations M49 Standard classification to which we adhered. Moreover, if we only consider applicants who disclose their sexual orientation via the signal of an LGB+ organisation, the regional difference is no longer significant (see Table A14–A15). In contrast with Flage (2020), we thus find no robust evidence that hiring discrimination based on sexual orientation is higher in Europe than in the Americas.

3.4.3. Period heterogeneity

There are diverging opinions about the extent to which discrimination has varied over time (Quillian et al., 2017). Recent meta-analytic evidence that relies on similar, causal evidence of hiring discrimination, however, suggested that there are few to no temporal changes in unequal treatment in hiring based on race, ethnicity, and national origin in the United States and the United Kingdom (Heath and Di Stasio, 2019; Quillian et al., 2017). Relying on data from recent correspondence experiments published between 2005 and 2020 (and conducted between 2002 and 2020), we reassess this evidence. At the same time, we also evaluate temporal heterogeneity in hiring discrimination by gender and motherhood status, age, religion, disability, and sexual orientation. Table 4 shows the results from the meta-regression including the time variable. A visual representation of this heterogeneity in hiring discrimination by discrimination ground and, in minor order, by region, sub-region, or country can be found in Appendix A (Figure A1–1 to Figure A1–16).

In contrast with the results of the meta-studies of Heath and Di Stasio (2019) and Quillian et al., (2017), we do initially find an overall decline in ethnic hiring discrimination. The correlation between (i) the weighted call-back ratios of the individual correspondence experiments related to race, ethnicity, and national origin and (ii) the years in which the respective experiments ended is negative and small but statistically significant, even after controlling for call-back classification and region effects ($r = -0.21$, $k = 143$; $\widehat{\beta} = 0.0173$, $p = 0.030$; see Table 4 and Figure A1–1).[28] This equates to an average increase in positive call-backs for ethnic minorities of 15.52 percentage points between 2006 ($\widehat{DR} = 0.6053$) and 2020 ($\widehat{DR} = 0.7605$).

The decline in ethnicity-based hiring discrimination is primarily driven by the moderate negative correlation related to European correspondence experiments ($r = -0.37$, $k = 94$; $\widehat{\beta} = 0.0267$, $p = 0.001$)—an average increase of 24.32 percentage points in positive call-backs for ethnic minorities between 2006 ($\widehat{DR} = 0.5886$) and 2020 ($\widehat{DR} = 0.8318$)—as opposed to studies conducted in the Americas, where no significant temporal change is observed ($r = 0.13$, $k = 38$; $p = 0.422$; see Table A18 and Figure A1–2).[29] Zooming in on the sub-regions, ethnic hiring discrimination seems to have been mainly in decline in Eastern Europe ($r = -0.84$, $k = 8$) and Western Europe ($r = -0.42$, $k = 48$; see Figure A1–3)—although the former finding is based on too few studies to make conclusive claims.

However, the temporal decrease in ethnic hiring discrimination in Europe does not occur to be robust. The decline becomes statistically insignificant when controlling for the considered treatment groups or when restricting the analysis to each country for which we have data separately (see Table A16, Table A18 and Figure A1–4). In other words, the choice of minority groups in correspondence experiments across European countries in combination with the timing of the experiments seems to have played a meaningful role in the declining figures of hiring discrimination based on race, ethnicity, and national origin. More specifically, in the 2005–2014 period, applicants of Northern, Eastern, and Western European or Western Asian origin received proportionately more attention and applicants of Arab/Maghrebi/Middle Eastern origin less than in the 2015–2020 period, while the former applicant groups face less hiring

[27] As a reminder, Table B1 in Appendix B includes details about the ages of the candidates in the treatment and control groups in the related correspondence experiments.

[28] The correlation coefficient was calculated as the weighted correlation between the majority/minority response ratios of the individual correspondence experiments and the year these experiments ended. Similarly, the regression coefficient was derived from the weighted least squares (WLS) model with said response ratios as the dependent variable and the year the experiments ended as the independent variable. We used the majority/minority response ratios because this allows us to interpret a negative (positive) correlation in terms of a decrease (increase) in hiring discrimination. Weights were derived from the meta-analytic random-effects model (see section 2.4.1).

[29] If any effect, there rather seems to be an upward trend in ethnic hiring discrimination in the Americas.

discrimination than the latter group (see Section 3.3.1). Regarding the remaining discrimination grounds (i.e. gender and motherhood status, age, religion, disability, and sexual orientation), we also find no structural, robust evidence for varying levels of hiring discrimination in recent years.[30] Taken together, our general impression is that there is limited change in hiring discrimination across discrimination grounds over time.

3.5. Publication bias

Appendix A (Table A11 and Figure A2–1 to Figure A2–16) contains useful information to evaluate publication bias. Publication bias is a problem related to the in- and exclusion of studies in a meta-analysis, potentially resulting in an over- or underestimation of the pooled estimates (Harrer et al., 2021). Based on the assessment of (i) funnel plot asymmetry, (ii) the related bias statistics, and (ii) statistical differences between the unadjusted and publication bias-adjusted estimates obtained from the 'limit' meta-analyses, we find that there is potential publication bias regarding the discrimination grounds of race, ethnicity, and national origin and sexual orientation—we find no structural evidence for publication bias associated with the other discrimination grounds.[31,32]

As explained in Section 2.4.3, we have attempted to limit the influence of publication bias to a minimum. Where appropriate and relevant, (i) the bias-adjusted estimates were reported instead of the unadjusted estimates, (ii) results were cross-checked with the outlier-adjusted pooled discrimination ratios, and (iii) discrimination ratios and meta-regression models were also calculated at the lower-level treatment groups (instead of only at the level of the discrimination ground), where there is generally less between-study heterogeneity and thus less (influence of) outliers. Overall, and unless reported otherwise in Sections 3.2–3.4, publication bias had a limited impact on the interpretability of the results.

4. Conclusion

In this meta-analysis, we extensively documented and synthesised the recent hiring discrimination literature grounded in the correspondence testing method—i.e. the reference method of measurement that allows for a causal interpretation of the empirical evidence of unequal treatment in hiring. Unique to our study is the focus on differences in hiring discrimination across discrimination grounds. More concretely, based on experiments from around the world, we quantified the level of hiring discrimination for ten grounds based on which unequal treatment is forbidden under United States federal or state law: (i) race, ethnicity, and national origin, (ii) gender and motherhood status, (iii) age, (iv) religion, (v) disability, (vi) sexual orientation, (vii) physical appearance, (viii) wealth, (ix) marital status, and (x) military service or affiliation. Moreover, we assessed the heterogeneity in hiring discrimination according to the classification of the call-back variable, the region linked to the correspondence experiment, and the related period. Our study provides scholars and policymakers with an extensive comparison based on hiring discrimination research from across the world. Knowing which and to what extent minority groups face labour market inaccessibility is invaluable in tackling this issue. In the following paragraphs, we first discuss the most important results of our meta-analysis, followed by the limitations of our research together with some suggestions for future research.

We observe four notable findings from our analyses. Our first observation relates to the results concerning hiring discrimination at the level of the discrimination ground. Historically, research efforts have focused heavily on examining hiring discrimination based on race, ethnicity, and national origin. This research commitment is not unjustified: applicants with salient racial or ethnic characteristics considerably different from those of the respective majority group(s) in a given country are significantly less likely to receive positive responses to their applications. Specifically, ethnic minority candidates on average receive nearly one-third fewer positive responses to their applications than their majority counterparts. However, it appears that the unequal treatment of applicants with disabilities, older applicants, and less physically attractive applicants is equally problematic. Applicants with disabilities or who have an odd physical appearance receive about two-fifths fewer positive responses on average, while the penalty for old(er) applicants is just above one-third. In addition, we found more modest evidence of hiring discrimination based on religion, wealth, and marital status. Diversity policies, such as outreach campaigns and diversity training, as well as other remedial measures, should also focus on these discrimination grounds. Our meta-analysis underlines that 'diversity in the labour market' should also have a diverse interpretation.

Second, levels of hiring discrimination against the specific minority groups within the set of examined discrimination grounds generally differ substantially. For example, candidates of Arab, Maghrebi, or Middle Eastern origin are severely discriminated against in the hiring process, facing an estimated average reduced chance of a positive response of about two-fifths. At the same time, there is only weak evidence of discrimination against (White) European minority applicants. Therefore, in the first place, measures to decrease hiring discrimination should be targeted at those minority groups who are penalised the most. In this respect, our meta-analysis offers an account of the severity of hiring discrimination against a multitude of minority groups.

Third, we found that there is more hiring discrimination against older applicants (vis-à-vis younger applicants) in Europe (i.e.

[30] Based on Figure A1-16 in Appendix A, one might believe that discrimination based on sexual orientation has declined between 2007 and 2013. However, controlling for call-back classification and region effects, no significant effect of period remains (see Table A14–A15). Moreover, there are issues with publication bias—as a result, the findings related to this form of hiring discrimination bear limited generalisability.

[31] As a reminder, following Harrer et al. (2021) and Sterne et al. (2011), publication bias was only assessed for discrimination grounds and treatment groups where $k \geq 10$ (see Table A11). Hence, we did not calculate estimates for the discrimination grounds physical appearance, wealth, military service or affiliation, and marital status.

[32] Looking at the funnel plot regarding the discrimination ground 'disability', there is, however, one clear outlier (see Figure A2-14).

Belgium, France, the United Kingdom, Spain, and Sweden; approximately 50% fewer positive call-backs on average) versus the United States (approximately 30% fewer positive call-backs on average). This finding is in line with the historic employment rates of 55- to 64-year-olds in the respective countries. Future studies could look into the specific mechanisms that drive these regional differences. European countries or institutions might consequently be able to learn from contextual or policy differences with the United States to determine possible counteracting measures.

Fourth, we observed little differences in hiring discrimination over time. Controlling for call-back classification and region effects, we initially found that hiring discrimination based on race, ethnicity, and national origin had decreased in European correspondence experiments. This decline is primarily driven by audit studies from Western European countries. However, when controlling for the minority groups considered in each experiment, the slope becomes statistically insignificant. The original finding also contrasts with our results at the country level: for each country for which there were sufficient observations to calculate a trend separately, we found no evidence for a significant decline in ethnic hiring discrimination. Moreover, we did not find evidence for structural temporal changes in unequal treatment in hiring related to the remaining discrimination grounds within the scope of this review. Overall, hiring discrimination remains a pervasive issue.

Notwithstanding the important contributions of our review, there are a few limitations concerning the research methods we applied. First of all, our research is based on a synthesis of only correspondence experiments, whereas some earlier meta-studies that focused on specific discrimination grounds have also included in-person audits to paint a broader picture of hiring discrimination (e.g. Quillian et al., 2017). However, as we argued in the introduction, in-person audits face a critical limitation. Behavioural differences between applicants, which are hard to control for, could have an undesirable influence on an employer's assessment in a selection context and therefore muddle the relationship between the individual characteristics of interest (e.g. national origin) and the hiring decision.

Moreover, to some extent, our meta-analysis might suffer from publication bias because we did not consider unpublished manuscripts or non-English research. Nonetheless, we statistically evaluated and attempted to control for said bias. In cases where we suspected publication bias, we reported the results with the necessary caveats but, in general, there were few. Especially the estimates regarding sexual orientation appeared to be influenced by publication bias. In addition, the number of included correspondence experiments regarding marital status, wealth, and military service or affiliation was too small to draw any convincing conclusions. Where the density of evidence is low, more experimentation will be needed before scholars can draw more conclusive inferences.

Finally, we did not explain most of the variability around the hiring discrimination estimates, in part because the scope of our review was already very broad but also because many relevant covariates were not retrievable at the study level across discrimination grounds. The variability that we did explain using meta-regression techniques could not be interpreted causally. For example, it is unclear what exactly drove the regional differences between various European countries and the United States in age discrimination in hiring. Discrepancies in legislation and the resulting social norms might have played a role but we cannot rule out alternative explanations. We advise future studies to investigate the drivers of this variability. A first strategy could be to directly assess context heterogeneity via correspondence experiments by examining the correlation between several vacancy, occupation, organisation, or sector characteristics and unequal treatment in hiring at once, eliminating alternative explanations (see e.g. Kline et al., 2021). A second approach could be to use vignette experiments, randomly assigning participants across context factors to observe how prejudice or different stigmas and stereotypes vary (see e.g. Van Borm and Baert, 2022). A third method that we can think of is to apply appropriate meta-regression techniques to sufficiently specific research problems regarding a more restricted selection of minority groups, assessing the relationship between the available study-level variables and hiring discrimination (see e.g. Quillian et al., 2017). This would enable scholars to more precisely attribute the uncovered variance of the pooled discrimination ratios to relevant factors beyond the contextual heterogeneity in terms of call-back classification, region, and period discussed in this review

Funding

This study was conducted in the context of the EdisTools project. EdisTools is funded by Research Foundation – Flanders (Strategic Basic Research, S004119N).

CRediT authorship contribution statement

Louis Lippens: Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Siel Vermeiren:** Investigation, Data curation, Writing – original draft. **Stijn Baert:** Conceptualization, Methodology, Data curation, Writing – original draft, Writing – review & editing, Supervision, Funding acquisition.

Declaration of Competing Interest

There are no relevant financial or non-financial competing interests.

Data Availability

The data used in this study are available at the following URL: https://doi.org/10.34740/kaggle/dsv/4142915.

Acknowledgements

We are grateful to the authors of a handful of the audit studies included in this review for providing us with missing data to supplement our dataset and with feedback on a prior version of this paper. Moreover, we are thankful to Brecht Neyt, the participants of the 26th Spring Meeting of Young Economists, and the participants of the 19th IMISCOE Annual Conference for their helpful comments and suggestions. Last, we want to thank three anonymous reviewers for their feedback, which greatly helped to improve this work

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.euroecorev.2022.104315.

References

Abel, G.J., Sander, N., 2014. Quantifying global international migration flows. Science 343 (6178), 1520–1522. https://doi.org/10.1126/science.1248676.
Adamovic, M., 2020. Analyzing discrimination in recruitment: a guide and best practices for resume studies. International Journal of Selection and Assessment 28 (4), 445–464. https://doi.org/10.1111/ijsa.12298.
Adamovic, M., 2022. When ethnic discrimination in recruitment is likely to occur and how to reduce it: applying a contingency perspective to review resume studies. Human Resource Management Review 32 (2), 100832. https://doi.org/10.1016/j.hrmr.2021.100832.
Adamovic, M., Leibbrandt, A., 2022. A large-scale field experiment on occupational gender segregation and hiring discrimination. Industrial Relations: A Journal of Economy and Society. https://doi.org/10.1111/irel.12318. Advance online publication.
Altonji, J.G., & Blank, R.M. (1999). Race and gender in the labor market. In O. C. Ashenfelter & D. Card (eds.), *Handbook of Labor Economics* (Vol. 3, pp. 3143–3259). Elsevier. https://doi.org/10.1016/S1573-4463(99)30039-0.
Altman, D.G., Bland, J.M., 2003. Statistics notes: interaction revisited: the difference between two estimates. BMJ 326 (7382), 219. https://doi.org/10.1136/bmj.326.7382.219.
Ameri, M., Schur, L., Adya, M., Bentley, F. S., McKay, P., & Kruse, D. (2018). The disability employment puzzle: A field experiment on employer hiring behavior. ILR Review, 71(2), 329364. https://doi.org/10.1177/0019793917717474.
Asali, M., Pignatti, N., Skhirtladze, S., 2018. Employment discrimination in a former Soviet Union republic: evidence from a field experiment. J Comp Econ 46 (4), 1294–1309. https://doi.org/10.1016/j.jce.2018.09.001.
Baert, S., 2014. Career lesbians. Getting hired for not having kids? Industrial Relations Journal 45 (6), 543–561. https://doi.org/10.1111/irj.12078.
Baert, S., 2018. Hiring discrimination: an overview of (almost) all correspondence experiments since 2005. In: Gaddis, S.M. (Ed.), Audit studies: Behind the Scenes With theory, method, and Nuance. Springer, pp. 63–77. https://doi.org/10.1007/978-3-319-71153-9_3.
Baert, S., 2021. The iceberg decomposition: a parsimonious way to map the health of labour markets. Econ Anal Policy 69, 350–365. https://doi.org/10.1016/j.eap.2020.12.012.
Baert, S., Norga, J., Thuy, Y., Van Hecke, M., 2016. Getting grey hairs in the labour market. An alternative experiment on age discrimination. J Econ Psychol 57, 86–101. https://doi.org/10.1016/j.joep.2016.10.002.
Balduzzi, S., Rücker, G., Schwarzer, G., 2019. How to perform a meta-analysis with R: a practical tutorial. Evidence Based Mental Health 22 (4), 153–160. https://doi.org/10.1136/ebmental-2019-300117.
Balestra, C., & Fleischer, L. (2018). Diversity statistics in the OECD: how do OECD countries collect data on ethnic, racial and indigenous identity? (Organisation for Economic Cooperation and Development [OECD] Statistics Working Papers No. 2018/09). Organisation for Economic Cooperation and Development. https://doi.org/10.1787/89bae654-en.
Bartkoski, T., Lynch, E., Witt, C., Rudolph, C., 2018. A meta-analysis of hiring discrimination against Muslims and Arabs. Personnel Assessment and Decisions 4 (2), 1–16. https://doi.org/10.25035/pad.2018.02.001.
Batinović, L., Howe, M., Sinclair, S., Carlsson, R., 2022. Ageism in hiring: a systematic review and meta-analysis of age discrimination. PsyArXiv. https://doi.org/10.31234/osf.io/sbzmv.
Beam, E.A., Hyman, J., Theoharides, C., 2020. The relative returns to education, experience, and attractiveness for young workers. Econ Dev Cult Change 68 (2), 391–428. https://doi.org/10.1086/701232.
Bertrand, M., Duflo, E., 2017. Field experiments on discrimination. In: Banerjee, A.V., Duflo, E. (Eds.), Handbook of Economic Field Experiments, 1st ed., pp. 309–393. https://doi.org/10.1016/bs.hefe.2016.08.004.
Bertrand, M., Mullainathan, S., 2004. Are Emily and Greg more employable than Lakisha and Jamal, A field experiment on labor market discrimination. American Economic Review 94 (4), 991–1013. https://doi.org/10.1257/0002828042002561.
Blinder, A.S., 1973. Wage discrimination: reduced form and structural estimates. J Hum Resour 8 (4), 436–455. https://doi.org/10.2307/144855.
Borenstein, M., Hedges, L.V., Higgins, J.P.T., Rothstein, H.R., 2009. Introduction to Meta-Analysis. John Wiley & Sons. https://doi.org/10.1002/9780470743386.
Borjas, G., 2020. Labor market discrimination. In: Borjas, G. (Ed.), Labor Economics, 8th ed. McGraw-Hill, pp. 299–340.
Briner, R.B., & Denyer, D. (2012). Systematic review and evidence synthesis as a practice and scholarship tool. Oxford Handbooks Online. https://doi.org/10.1093/oxfordhb/9780199763986.013.0007.
Boystock v. Clayton County, 590 U.S. __ (U.S. Sup. Ct. 2020). https://www.supremecourt.gov/opinions/19pdf/17-1618_hfci.pdf.
Carlsson, M., Eriksson, S., 2019. Age discrimination in hiring decisions: evidence from a field experiment in the labor market. Labour Econ 59, 173–183. https://doi.org/10.1016/j.labeco.2019.03.002.
Cochran, W.G., 1954. Some methods for strengthening the common χ^2 tests. Biometrics 10 (4), 417–451. https://doi.org/10.2307/3001616.
Cortina, C., Rodríguez, J., González, M.J., 2021. Mind the job: the role of occupational characteristics in explaining gender discrimination. Soc Indic Res 156 (1), 91–110. https://doi.org/10.1007/s11205-021-02646-2.
Darolia, R., Koedel, C., Martorell, P., Wilson, K., Perez-Arce, F., 2016. Race and gender effects on employer interest in job applicants: new evidence from a resume field experiment. Appl Econ Lett 23 (12), 853–856. https://doi.org/10.1080/13504851.2015.1114571.
Derous, E., Ryan, A.M., 2019. When your resume is (not) turning you down: modelling ethnic bias in resume screening. Human Resource Management Journal 29 (2), 113–130. https://doi.org/10.1111/1748-8583.12217.
Di Stasio, V., Lancee, B., Veit, S., Yemane, R., 2021. Muslim by default or religious discrimination, Results from a cross-national field experiment on hiring discrimination. J Ethn Migr Stud 47 (6), 1305–1326. https://doi.org/10.1080/1369183x.2019.1622826.
Drydakis, N., 2009. Sexual orientation discrimination in the labour market. Labour Econ 16 (4), 364–372. https://doi.org/10.1016/j.labeco.2008.12.003.
Drydakis, N., 2014. Sexual orientation discrimination in the Cypriot labour market: distastes or uncertainty? Int J Manpow 35 (5), 720–744. https://doi.org/10.1108/ijm-02-2012-0026.

Drydakis, N., 2017. Measuring labour differences between natives, non-natives, and natives with an ethnic-minority background. Econ Lett 161, 27–30. https://doi.org/10.1016/j.econlet.2017.08.031.
Drydakis, N., 2022. Sexual orientation and earnings: a meta-analysis 2012–2020. J Popul Econ 35 (2), 409–440. https://doi.org/10.1007/s00148-021-00862-1.
European Commission. (2021). Guidance note on the collection and use of equality data based on racial or ethnic origin. Publications Office of the European Union. https://ec.europa.eu/info/sites/default/files/guidance_note_on_the_collection_and_use_of_equality_data_based_on_racial_or_ethnic_origin.pdf.
Flage, A., 2020. Discrimination against gays and lesbians in hiring decisions: a meta-analysis. Int J Manpow 41 (6), 671–691. https://doi.org/10.1108/ijm-08-2018-0239.
Gaddis, S.M., 2015. Discrimination in the credential society: an audit study of race and college selectivity in the labor market. Social Forces 93 (4), 1451–1479. https://doi.org/10.1093/sf/sou111.
Gaddis, S.M., 2017. Racial/Ethnic perceptions from Hispanic names: selecting names to test for discrimination. Socius: Sociological Research for a Dynamic World 3, 1–11. https://doi.org/10.1177/2378023117737193.
Gaddis, S.M., 2018. An Introduction to audit studies in the social sciences. In: Gaddis, S.M. (Ed.), Audit studies: Behind the Scenes With theory, method, and Nuance. Springer, pp. 3–44. https://doi.org/10.1007/978-3-319-71153-9_1.
Gaddis, S.M., Larsen, E., Crabtree, C., & Holbein, J. (2021). Discrimination against Black and Hispanic Americans is highest in hiring and housing contexts: a meta-analysis of correspondence audits. (SSRN Electronic Journal Working Paper No. 3975770). University of Chicago, Becker Friedman Institute for Economics. https://doi.org/10.2139/ssrn.3975770.
Ganty, S. & Benito-Sanchez, J.C. (2021). Expanding the list of protected grounds within anti-discrimination law in the EU. Equinet: European Network of Equality Bodies. https://equineteurope.org/wp-content/uploads/2022/03/Expanding-the-List-of-Grounds-in-Non-discrimination-Law_Equinet-Report.pdf.
Guul, T.S., Villadsen, A.R., Wulff, J.N., 2019. Does good performance reduce bad behavior, Antecedents of ethnic employment discrimination in public organizations. Public Adm Rev 79 (5), 666–674. https://doi.org/10.1111/puar.13094.
Harrer, M., Cuijpers, P., Furukawa, T.A., & Ebert, D.D. (2019). dmetar: Companion R package for the guide 'Doing meta-analysis in R' (Version 0.0.9000) [Computer software]. https://dmetar.protectlab.org.
Harrer, M., Cuijpers, P., Furukawa, T.A., Ebert, D.D., 2021. Doing Meta-Analysis With R: A hands-On Guide, 1st ed. Chapman and Hall/CRC. https://doi.org/10.1201/9781003107347.
Havránek, T., Stanley, T.D., Doucouliagos, H., Bom, P., Geyer-Klingeberg, J., Iwasaki, I., Reed, W.R., Rost, K., Aert, R.C.M., 2020. Reporting guidelines for meta-analysis in economics. J Econ Surv 34 (3), 469–475. https://doi.org/10.1111/joes.12363.
Heath, A.F., Di Stasio, V., 2019. Racial discrimination in Britain, 1969–2017: a meta-analysis of field experiments on racial discrimination in the British labour market. Br J Sociol 70 (5), 1774–1798. https://doi.org/10.1111/1468-4446.12676.
Higgins, J.P.T., Thompson, S.G., 2002. Quantifying heterogeneity in a meta-analysis. Stat Med 21 (11), 1539–1558. https://doi.org/10.1002/sim.1186.
Higgins, J.P.T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M.J., Welch, V.A., 2019. Cochrane Handbook For Systematic Reviews of Interventions, 1st ed. Wiley. https://doi.org/10.1002/9781119536604.
Hoaglin, D.C., 2016. Misunderstandings about Q and 'Cochran's Q test' in meta-analysis. Stat Med 35 (4), 485–495. https://doi.org/10.1002/sim.6632.
Jacquemet, N., Yannelis, C., 2012. Indiscriminate discrimination: a correspondence test for ethnic homophily in the Chicago labor market. Labour Econ 19 (6), 824–832. https://doi.org/10.1016/j.labeco.2012.08.004.
Jowell, R., Prescott-Clarke, P., 1970. Racial discrimination and white-collar workers in Britain. Race 11 (4), 397–417. https://doi.org/10.1177/030639687001100401.
Kitagawa, E.M., 1955. Components of a difference between two rates. J Am Stat Assoc 50 (272), 1168. https://doi.org/10.2307/2281213.
Kline, P., Rose, E., & Walters, C. (2021). Systemic discrimination among large US employers. (SSRN Electronic Journal Working Paper No. 2021–94). University of Chicago, Becker Friedman Institute for Economics. https://doi.org/10.2139/ssrn.3898669.
Knapp, G., Hartung, J., 2003. Improved tests for a random effects meta-regression with a single covariate. Stat Med 22 (17), 2693–2710. https://doi.org/10.1002/sim.1482.
Lahey, J.N., 2010. International comparison of age discrimination laws. Res Aging 32 (6), 679–697. https://doi.org/10.1177/0164027510379348.
Lang, K., Kahn-Lang-Spitzer, A., 2020. Race discrimination: an economic perspective. Journal of Economic Perspectives 34 (2), 68–89. https://doi.org/10.1257/jep.34.2.68.
Langan, D., Higgins, J.P.T., Jackson, D., Bowden, J., Veroniki, A.A., Kontopantelis, E., Viechtbauer, W., Simmonds, M., 2019. A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. Res Synth Methods 10 (1), 83–98. https://doi.org/10.1002/jrsm.1316.
Larsen, E.N., Di Stasio, V., 2021. Pakistani in the UK and Norway: different contexts, similar disadvantage. Results from a comparative field experiment on hiring discrimination. J Ethn Migr Stud 47 (6), 1201–1221. https://doi.org/10.1080/1369183x.2019.1622777.
Lippens, L., Baert, S., Ghekiere, A., Verhaeghe, P.-P., Derous, E., 2022. Is labour market discrimination against ethnic minorities better explained by taste or statistics? A systematic review of the empirical evidence. J Ethn Migr Stud. https://doi.org/10.1080/1369183X.2022.2050191. Advance online publication.
Lippens, L., Vermeiren, S., & Baert, S. (2021). The state of hiring discrimination: a meta-analysis of (almost) all recent correspondence experiments (Institut zur Zukunft der Arbeit [IZA] Discussion Papers No. 14966). IZA Institute of Labor Economics. https://www.iza.org/publications/dp/14966/the-state-of-hiring-discrimination-a-meta-analysis-of-almost-all-recent-correspondence-experiments.
Mantel, N., Haenszel, W., 1959. Statistical aspects of the analysis of data from retrospective studies of disease. J. Natl. Cancer Inst. 22 (4), 719–748. https://doi.org/10.1093/jnci/22.4.719.
Maurer-Fazio, M., 2013. Ethnic discrimination in China's internet job board labor market. IZA Journal of Migration 1 (12). https://doi.org/10.1186/2193-9039-1-12.
Morning, A., 2008. Ethnic classification in global perspective: a cross-national survey of the 2000 census round. Popul Res Policy Rev 27 (2), 239–272. https://doi.org/10.1007/s11113-007-9062-5.
Neumark, D., 2018. Experimental research on labor market discrimination. J Econ Lit 56 (3), 799–866. https://doi.org/10.1257/jel.20161309.
Neumark, D., Burn, I., Button, P., Chehras, N., 2019. Do state laws protecting older workers from discrimination reduce age discrimination in hiring? Evidence from a field experiment. The Journal of Law and Economics 62 (2), 373–402. https://doi.org/10.1086/704008.
Oaxaca, R., 1973. Male-female wage differentials in urban labor markets. Int Econ Rev (Philadelphia) 14 (3), 693–709. https://doi.org/10.2307/2525981.
Organisation for Economic Cooperation and Development. (2020a). All hands in? Making diversity work for all. https://doi.org/10.1787/efb14583-en.
Organisation for Economic Cooperation and Development. (2020b). International migration outlook 2020. https://doi.org/10.1787/0c0cc42a-en.
Organisation for Economic Cooperation and Development. (2021). Employment rate by age group [Data set]. https://doi.org/10.1787/084f32c7-en.
Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., Chou, R., Glanville, J., Grimshaw, J.M., Hróbjartsson, A., Lalu, M.M., Li, T., Loder, E.W., Mayo-Wilson, E., McDonald, S., Moher, D., 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 372, n71. https://doi.org/10.1136/bmj.n71.
Page, M.J., Sterne, J.A.C., Higgins, J.P.T., Egger, M., 2020. Investigating and dealing with publication bias and other reporting biases in meta-analyses of health research: a review. Res Synth Methods 12 (2), 248–259. https://doi.org/10.1002/jrsm.1468.
Pager, D., 2016. Are firms that discriminate more likely to go out of business. Sociol Sci 3, 849–859. https://doi.org/10.15195/v3.a36.
Patacchini, E., Ragusa, G., Zenou, Y., 2015. Unexplored dimensions of discrimination in Europe: homosexuality and physical appearance. J Popul Econ 28 (4), 1045–1073. https://doi.org/10.1007/s00148-014-0533-9.
Peters, J.L., 2006. Comparison of two methods to detect publication bias in meta-analysis. JAMA 295 (6), 676–680. https://doi.org/10.1001/jama.295.6.676.
Quillian, L., Midtbøen, A.H., 2021. Comparative perspectives on racial discrimination in hiring: the rise of field experiments. Annu Rev Sociol 47 (1), 391–415. https://doi.org/10.1146/annurev-soc-090420-035144.
Quillian, L., Heath, A., Pager, D., Midtbøen, A., Fleischmann, F., Hexel, O., 2019. Do some countries discriminate more than others? Evidence from 97 field experiments of racial discrimination in hiring. Sociol Sci 6, 467–496. https://doi.org/10.15195/v6.a18.

Quillian, L., Lee, J.J., Oliver, M., 2020. Evidence from field experiments in hiring shows substantial additional racial discrimination after the callback. Social Forces 99 (2), 732–759. https://doi.org/10.1093/sf/soaa026.
Quillian, L., Pager, D., Hexel, O., Midtbøen, A.H., 2017. Meta-analysis of field experiments shows no change in racial discrimination in hiring over time. Proceedings of the National Academy of Sciences 114 (41), 10870–10875. https://doi.org/10.1073/pnas.1706255114.
Riach, P.A., 2015. A field experiment investigating age discrimination in four European labour markets. International Review of Applied Economics 29 (5), 608–619. https://doi.org/10.1080/02692171.2015.1021667.
Rich, J. (2014). What do field experiments of discrimination in markets tell us? A meta-analysis of studies conducted since 2000 (IZA Discussion Papers No. 8584). IZA Institute of Labor Economics. https://www.iza.org/publications/dp/8584/what-do-field-experiments-of-discrimination-in-markets-tell-us-a-meta-analysis-of-studies-conducted-since-2000.
Richardson, W.S., Wilson, M.C., Nishikawa, J., Hayward, R.S.A., 1995. The well-built clinical question: a key to evidence-based decisions. ACP J. Club 123 (3), A12. https://doi.org/10.7326/acpjc-1995-123-3-a12.
Rücker, G., Schwarzer, G., Carpenter, J.R., Binder, H., Schumacher, M., 2011. Treatment-effect estimates adjusted for small-study effects via a limit meta-analysis. Biostatistics 12 (1), 122–142. https://doi.org/10.1093/biostatistics/kxq046.
Rücker, G., Schwarzer, G., Carpenter, J.R., Schumacher, M., 2008. Undue reliance on *I2* in assessing heterogeneity may mislead. BMC Med Res Methodol 8 (1), 79. https://doi.org/10.1186/1471-2288-8-79.
Schwarzer, G., Carpenter, J.R., Rücker, G, 2015. Heterogeneity and meta-regression. In: Schwarzer, G., Carpenter, J.R., Rücker, G. (Eds.), Meta-analysis With R. Springer, pp. 85–104. https://doi.org/10.1007/978-3-319-21416-0_4.
Schwarzer, G., Carpenter, J.R., & Rücker, G. (2020). Metasens: advanced statistical methods to model and adjust for bias in meta-analysis (Version 0.6–0) [Computer software]. https://CRAN.R-project.org/package=metasens.
Sterne, J.A.C., Sutton, A.J., Ioannidis, J.P.A., Terrin, N., Jones, D.R., Lau, J., Carpenter, J., Rucker, G., Harbord, R.M., Schmid, C.H., Tetzlaff, J., Deeks, J.J., Peters, J., Macaskill, P., Schwarzer, G., Duval, S., Altman, D.G., Moher, D., Higgins, J.P.T., 2011. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ 343, d4002. https://doi.org/10.1136/bmj.d4002.
Stone, A., Wright, T., 2013. When your face doesn't fit: employment discrimination against people with facial disfigurements. J Appl Soc Psychol 43 (3), 515–526. https://doi.org/10.1111/j.1559-1816.2013.01032.x.
Thijssen, L., Coenders, M., Lancee, B., 2021a. Ethnic discrimination in the Dutch labor market: differences between ethnic minority groups and the role of personal information about job applicants—evidence from a field experiment. J Int Migr Integr 22 (3), 1125–1150. https://doi.org/10.1007/s12134-020-00795-w.
Thijssen, L., van Tubergen, F., Coenders, M., Hellpap, R., Jak, S., 2021b. Discrimination of Black and Muslim minority groups in western societies: evidence from a meta-analysis of field experiments. International Migration Review. https://doi.org/10.1177/01979183211045044. Advance online publication.
Thomas, K., 2018. The labor market value of taste: an experimental study of class bias in US employment. Sociol Sci 5, 562–595. https://doi.org/10.15195/v5.a24.
Tilcsik, A., 2011. Pride and prejudice: employment discrimination against openly gay men in the United States. American Journal of Sociology 117 (2), 586–626. https://doi.org/10.1086/661653.
United Nations. (2021). United Nations standard country codes (Series M: miscellaneous Statistical Papers No. 49). https://unstats.un.org/unsd/methodology/m49/.
Van Borm, H., & Baert, S. (2022). Diving in the minds of recruiters: what triggers gender stereotypes in hiring? (IZA Discussion Papers No. 15261). IZA Institute of Labor Economics. https://www.iza.org/publications/dp/15261/diving-in-the-minds-of-recruiters-what-triggers-gender-stereotypes-in-hiring.
Verhaeghe, P.-P., 2022. Correspondence studies. In: Zimmermann, K.F. (Ed.), Handbook of Labor, Human Resources and Population Economics. Springer, Cham. https://doi.org/10.1007/978-3-319-57365-6_306-1.
Veroniki, A.A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J.P.T., Langan, D., Salanti, G., 2015. Methods to estimate the between-study variance and its uncertainty in meta-analysis. Res Synth Methods 7 (1), 55–79. https://doi.org/10.1002/jrsm.1164.
Viechtbauer, W., 2010. Conducting meta-analyses in R with the metafor package. J Stat Softw 36 (3), 1–48. https://doi.org/10.18637/jss.v036.i03.
Viechtbauer, W., López-López, J.A., Sánchez-Meca, J., Marín-Martínez, F., 2015. A comparison of procedures to test for moderators in mixed-effects meta-regression models. Psychol Methods 20 (3), 360–374. https://doi.org/10.1037/met0000023.
Yavorsky, J.E., 2019. Uneven patterns of inequality: an audit analysis of hiring-related practices by gendered and classed contexts. Social Forces 98 (2), 461–492. https://doi.org/10.1093/sf/soy123.
Yemane, R., Fernández-Reino, M., 2021. Latinos in the United States and in Spain: the impact of ethnic group stereotypes on labour market outcomes. J Ethn Migr Stud 47 (6), 1240–1260. https://doi.org/10.1080/1369183x.2019.1622806.
Zschirnt, E., Ruedin, D., 2016. Ethnic discrimination in hiring decisions: a meta-analysis of correspondence tests, 1990–2015. J Ethn Migr Stud 42 (7), 1115–1134. https://doi.org/10.1080/1369183X.2015.1133279.

E-lit-013: Detecting Discrimination

James J. Heckman | Journal of Economic Perspectives 12(2): 101-116 | 1998

https://web.archive.org/web/20190416193944id_/https://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.12.2.101

About this source

Heckman makes three claims that bear directly on the section's hypothesis. First, the audit method identifies discrimination only under an assumption nobody has verified: that the mean of the productivity characteristics the audit does not control is equal across race groups. Second, and more damaging, if firms hire by a threshold rule — productivity must exceed a cutoff — then equal means are insufficient; if the distributions of unobserved characteristics differ in dispersion, hiring rates will differ even with an identical cutoff for both groups, and the sign of the measured gap flips with the level at which the auditors' observed qualifications were set. Standardizing high favors the less dispersed group; standardizing low favors the more dispersed one, so the method can produce apparent discrimination, apparent reverse discrimination, or a null in an economy with no discrimination at all, and can equally hide real discrimination when blacks face a higher cutoff. Third, matching pairs tightly can make matters worse, not better: it strips out the common measured components and leaves the hiring gap driven by idiosyncratic unobserved factors, so the bias can exceed that from comparing random pairs. Reading the audits themselves, he stresses the symmetry rather than the gap — black and white auditors were treated identically in about 86 percent of Chicago and Denver attempts and 75.1 percent in Washington, with whites favored in 9.6 percent (Chicago), 19.1 percent (Washington) and 4.8 percent (Denver), and blacks favored in 4.5, 5.8 and 8.3 percent — and concludes the pattern is 'intrinsically ambiguous'. He separately argues that individual-firm discrimination need not imply market discrimination, since minorities purposively sort away from the worst firms, so a well-designed audit can find many discriminating firms while the marginal wage effect is zero; and that most of the residual 12 to 15 percent black-white wage gap reflects pre-market skill differences, of which 1 to 2 percentage points come from the GED miscoding alone.

Bound as the original PDF. Retrieved 2026-09-10. sha256 53e5958984502173…

Bound into an Economic Briefing · by Claude, may contain mistakes

Full document begins on the next page.

E-lit-013: https://web.archive.org/web/20190416193944id_/https://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.12.2.101

Journal of Economic Perspectives—Volume 12, Number 2—Spring 1998—Pages 101–116

Detecting Discrimination

James J. Heckman

In the current atmosphere of race relations in America, the authors of the three main papers presented in this symposium are like persons crying ''fire'' in a crowded theater. They apparently vindicate the point of view that American society is riddled with racism and that discrimination by employers may account for much of the well-documented economic disparity between blacks and whites. In my judgement, this conclusion is not sustained by a careful reading of the evidence.

In this article, I make three major points. First, I want to distinguish market discrimination from the discrimination encountered by a randomly selected person or pair of persons at a randomly selected firm as identified from audit studies.

Second, I consider the evidence presented by the authors in the symposium, focusing for brevity and specificity on labor markets. It is far less decisive on the issue of market discrimination than it is claimed to be. Disparity in market outcomes does not prove discrimination in the market. A careful reading of the entire body of available evidence confirms that most of the disparity in earnings between blacks and whites in the labor market of the 1990s is due to the differences in skills they bring to the market, and not to discrimination within the labor market. This interpretation of the evidence has important consequences for social policy. While undoubtedly there are still employers and employees with discriminatory intentions, labor market discrimination is no longer a first-order quantitative problem in American society. At this time, the goal of achieving black economic progress is better served by policies that promote skill formation, like improving family environments,

■ *James J. Heckman is Henry Schultz Distinguished Service Professor of Economics, University of Chicago, Chicago, Illinois, and Senior Fellow, American Bar Foundation, Chicago, Illinois.*

schools and neighborhoods, not by strengthening the content and enforcement of civil rights laws—the solution to the problem of an earlier era.

Third, I want to examine the logic and limitations of the audit pair method. All of the papers in this symposium use evidence from this version of pair matching. However, the evidence acquired from it is less compelling than is often assumed. Inferences from such studies are quite fragile to alternative assumptions about unobservable variables and the way labor markets work. The audit method can find discrimination when in fact none exists; it can also disguise discrimination when it is present. These findings are especially troubling because the Equal Employment Opportunity Commission has recently authorized the use of audit pair methods to detect discrimination in labor markets (Seelye, 1997).

Discrimination: Definition and Measurement

The authors of these papers focus on the question of whether society is color blind, not on the specific question of whether there is market discrimination in realized transactions. But discrimination at the individual level is different from discrimination at the group level, although these concepts are often confused in the literature on the economics of discrimination.

At the level of a potential worker or credit applicant dealing with a firm, racial discrimination is said to arise if an otherwise identical person is treated differently by virtue of that person's race or gender, and race and gender by themselves have no direct effect on productivity. Discrimination is a causal effect defined by a hypothetical *ceteris paribus* conceptual experiment—varying race but keeping all else constant. Audit studies attempt to identify racial and gender discrimination so defined for the set of firms sampled by the auditors by approximating the *ceteris paribus* condition.

It was Becker's (1957) insight to observe that finding a discriminatory effect of race or gender at a randomly selected firm does not provide an accurate measure of the discrimination that takes place in the market as a whole. At the level of the market, the causal effect of race is defined by the marginal firm or set of firms with which the marginal minority member deals. The impact of market discrimination is not determined by the most discriminatory participants in the market, or even by the average level of discrimination among firms, but rather by the level of discrimination at the firms where ethnic minorities or women actually end up buying, working and borrowing. It is at the margin that economic values are set. This point is largely ignored in the papers in this symposium.

This confusion between individual firm and market discrimination arises in particular in the audit studies. A well-designed audit study could uncover many individual firms that discriminate, while at the same time the marginal effect of discrimination on the wages of employed workers could be zero. This helps to explain the gap between audit-based estimates of discrimination and estimates

based on actual purchase prices that are discussed by Yinger. In fact, the audit evidence reported in all three papers in this symposium is entirely consistent with little or no market discrimination at the margin. Purposive sorting within markets eliminates the worst forms of discrimination. There may be evil lurking in the hearts of firms that is never manifest in consummated market transactions.

Estimating the extent and degree of discrimination, whether at the individual or the market level, is a difficult matter. In the labor market, for example, a worker's productivity is rarely observed directly, so the analyst must instead use available data as a proxy in controlling for the relevant productivity characteristics. The major controversies arise over whether relevant omitted characteristics differ between races, and between genders, and whether certain included characteristics systematically capture productivity differences or instead are a proxy for race or gender.

How Substantial is Labor Market Discrimination Against Blacks?

In their paper in this symposium, Darity and Mason present a bleak picture of the labor market position of African-Americans in which market discrimination is ubiquitous. They present a quantitative estimate of the magnitude of estimated discrimination: 12 to 15 percent in both 1980 and 1990 using standard regressions fit on Current Population Survey and Census data. Similar regressions show that the black/white wage gap has diminished sharply over the last half century. Comparable estimates for 1940 show a black/white wage gap ranging from 30 percentage points, for men age 25–34, to 42 percentage points, men age 55–64. In 1960, the corresponding numbers would have been 21 percent and 32 percent, for the same two age groups; in 1970, 18 and 25 percent (U.S. Commission on Civil Rights, 1986, Table 6.1, p. 191). The progress was greatest in Southern states where a blatantly discriminatory system was successfully challenged by an external legal intervention (Donohue and Heckman, 1991; Heckman, 1990).

How should the residual wage gap be interpreted? As is typical of much of the literature on measuring racial wage gaps, Darity and Mason never precisely define the concept of discrimination they use. As is also typical of this literature, the phrase ''human capital variable'' is thrown around without a clear operational definition. The implicit definition of these terms varies across the studies they discuss. In practice, human capital in these studies has come to mean education and various combinations of age and education, based on the available Census and Current Population Survey (CPS) data. However, there is a staggering gap between the list of productivity characteristics available to economic analysts in standard data sources and what is available to personnel departments of firms. Regressions based on the Census and/or CPS data can typically explain 20 to 30 percent of the variation in wages. However, regressions based on personnel data can explain a substantially higher share of the variation in wages; 60–80 percent in professional labor markets (for example, see Abowd and Killingsworth, 1983). It is not idle speculation to claim

that the standard data sets used to estimate discrimination omit many relevant characteristics actually used by firms in their hiring and promotion decisions. Nor is it idle speculation to conjecture that disparity in family, neighborhood and schooling environments may account for systematic differences in unmeasured characteristics between race groups.

Consider just one well-documented source of discrepancy between Census variables and the productivity concepts that they proxy: the measurement of high school credentials. The standard Census and CPS data sources equate recipients of a General Equivalence Degree, or GED, with high school graduates. However, black high school certificate holders are much more likely than whites to receive GEDs (Cameron and Heckman, 1993), and a substantial portion of the widely trumpeted ''convergence'' in measured black educational attainment has come through GED certification. Thus, in 1987 in the NLSY data that Darity and Mason discuss, and Neal and Johnson (1996) analyze, 79 percent of black males age 25 were high school certified, and 14 percent of the credential holders were GED recipients. Among white males, 88 percent were high school certified, and only 8 percent of the white credential holders were GED certified. Given the evidence from Cameron and Heckman that GED recipients earn the same as high school dropouts, it is plausible that standard Census-based studies that use high school credentials to control for ''education'' will find that the wages of black high school ''graduates'' are lower than those of whites.

Most of the empirical literature cited by Darity and Mason takes Census variables literally and ignores these issues. The GED factor alone accounts for 1–2 percentage points of the current 12–15 percent black-white hourly wage gap. An enormous body of solid evidence on inferior inner city schools and poor neighborhoods makes the ritual of the measurement of ''discrimination'' using the unadjusted Census or Current Population Survey data a questionable exercise.

Darity and Mason bolster their case for rampant discrimination by appealing to audit pair evidence. They do not point out that audit pair studies have primarily been conducted for hiring in entry level jobs in certain low skill occupations using overqualified college students during summer vacations. They do not sample subsequent promotion decisions. They fail to point out that the audits undersample the main avenues through which youth get jobs, since only job openings advertised in newspapers are audited, and not jobs found through networks and friends (Heckman and Siegelman, 1993, pp. 213–215). Auditors are sometimes instructed on the ''problem of discrimination in American society'' prior to sampling firms, so they may have been coached to find what the audit agencies wanted to find. I have already noted that audit evidence does not translate into actual employment experiences and wages obtained by actors who purposively search markets.

Putting these objections to the side, what do the audits actually show for this unrepresentative snapshot of the American labor market? Table 1 presents evidence

Table 1
Outcomes From Major Audit Studies For Blacks
(outcome: get job or not)

Number of Audits	*Pair*	*(a) Both Get Job*	*(b) Neither Gets a Job*	*Equal Treatment a + b*	*White Yes, Black No*	*White No, Black Yes*
Chicago*						
35	1	(5) 14.3%	(23) 65.7%	80.0%	(5) 14.3%	(2) 5.7%
40	2	(5) 12.5%	(25) 62.5%	75.0%	(4) 10.0%	(6) 15.0%
44	3	(3) 6.8%	(37) 84.1%	90.9%	(3) 6.8%	(1) 2.3%
36	4	(6) 16.7%	(24) 66.7%	83.4%	(6) 16.7%	(0) 0%
42	5	(3) 7.1%	(38) 90.5%	97.6%	(1) 2.4%	(0) 0%
197	Total	(22) 11.2%	(147) 74.6%	85.8%	(19) 9.6%	(9) 4.5%
Washington*						
46	1	(5) 10.9%	(26) 56.5%	67.4%	(12) 26.1%	(3) 6.5%
54	2	(11) 20.4%	(31) 57.4%	77.8%	(9) 16.7%	(3) 5.6%
62	3	(11) 17.7%	(36) 58.1%	75.8%	(11) 17.7%	(4) 6.5%
37	4	(6) 16.2%	(22) 59.5%	75.7%	(7) 18.9%	(2) 5.4%
42	5	(7) 16.7%	(26) 61.9%	77.6%	(7) 16.7%	(2) 4.8%
241	Total	(40) 16.6%	(141) 58.5%	75.1%	(46) 19.1%	(14) 5.8%
Denver**						
18	1	(2) 11.1%	(11) 61.1%	72.1%	(5) 27.8%	(0) 0.0%
53	2	(2) 3.8%	(41) 77.4%	81.2%	(0) 0.0%	(10) 18.9%
33	3	(7) 21.2%	(25) 75.8%	97.0%	(1) 3.0%	(0) 0.0%
15	4	(9) 60.0%	(3) 20.0%	80.0%	(2) 6.7%	(2) 13.3%
26	9	(3) 11.5%	(23) 88.5%	100.0%	(0) 0.0%	(0) 0.0%
145	Total	(23) 15.8%	(103) 71.1%	86.9%	(7) 4.8%	(12) 8.3%

Note: Results are percentages; figures in parentheses are the relevant number of audits.
Sources: Heckman and Siegelman (1993).
* This study was conducted by the Urban Institute.
** Denver pair numbers are for both black and Hispanic audits. For the sake of brevity, I only consider the black audits. The Denver study was not conducted by the Urban Institute but it was conducted to conform to Urban Institute practice.

from three major audits in Washington, D.C., Chicago and Denver. The most remarkable feature of this evidence is the a + b column which records the percentage of audit attempts where black and white auditors were treated symmetrically (both got a job; neither got a job). In Chicago and Denver this happened about 86 percent of the time. The evidence of disparity in hiring presented in the last two columns of the table suggests only a slight preference for whites over minorities; in several pairs, minorities are favored. Only a zealot can see evidence in these data of pervasive discrimination in the U.S. labor market. And, as I will show in the next section, even this evidence on disparity has to be taken with a grain of salt, because it is based on the implicit assumption that the distribution of unobserved productivity is the same in both race groups.

Darity and Mason go on to dismiss the research of Neal and Johnson (1996) who analyze a sample of males who took an achievement or ability test in their early teens—specifically, the Armed Forces Qualifications Test (AFQT)—and ask how much of the gap in black-white wages measured a decade or so after the test was taken can be explained by the differences in the test scores.[1] It is remarkable and important that this early ''premarket'' measure of ability plays such a strong role in explaining wages measured a decade after the test is taken. This is as true for studies of white outcomes taken in isolation as it is for black-white comparisons. Their findings are important for interpreting the sources of black-white disparity in labor market outcomes.

The goal of Neal and Johnson is not to estimate the racial gap by holding other characteristics constant in the sense of a linear regression. Their estimated race effect is obtained from an equation with essentially only ability and race in it. Ability is allowed to operate both as a direct effect, as in a standard hedonic regression with productivity characteristics included, and as an indirect effect, operating through subsequent schooling, work experience and occupational choices. Therefore, they strip out of their wage regression all of the schooling, occupation and other post-adolescent choice variables used in predecessor studies and commonly employed in ''human capital discrimination studies.'' In so doing, they bypass the problem of determining the endogeneity of the included control variables.

The Neal-Johnson story is not about genetic determination. They demonstrate that schooling and environment can affect their measured test score. A huge body of evidence, to which the Neal-Johnson study contributes, documents that human abilities and motivations are formed early and have a decisive effect on lifetime outcomes; the evidence is summarized in Heckman (1995) and in Heckman, Lochner, Taber and Smith (1997). Not only is early ability an important predictor of later success for blacks or whites, it can be manipulated. Early interventions are far more effective than late ones because early skills and motivation beget later skills and motivation. As Heckman, Lochner, Taber and Smith document, however, successful early interventions can be quite costly.

The objections raised by Darity and Mason against the Neal-Johnson study are largely specious. For example, Rodgers and Spriggs (1996) miss the point of the Neal-Johnson article by ''adjusting'' the test score by a later variable, such as schooling. But ability is known to be an important determinant of schooling (Cawley, Heckman and Vtylacil, 1998), so it should be no surprise that ''adjusting'' the score for later schooling eliminates an important component of ability and that adjusted scores play a much weaker role in explaining black-white differentials.[2]

[1] Specifically, Darity and Mason write: ''This effort has uncovered one variable in one data set which, if inserted in an earnings regression, produces the outcome that nearly all of the black male-white male wage gap is explained by human capital and none by labor market discrimination.''

[2] The Rodgers and Spriggs comment (1997) on Neal-Johnson raises other red herrings. Their confused discussion of endogeneity of AFQT, and their ''solution'' to the problem end up with an ''adjusted'' AFQT measure that is poorly correlated with the measured AFQT, and so is a poor proxy for black ability.

Only one point raised by Darity and Mason concerning Neal and Johnson is potentially valid—and this is a point made by Neal and Johnson in their original article. Black achievement scores may be lower than white scores not because of the inferior environments encountered by many poor blacks, but because of expectations of discrimination in the market. If black children and their parents face a world in which they receive lower rewards for obtaining skills, they will invest less if they face the same tuition costs as whites. Poor performance in schools and low achievement test scores may thus be a proxy for discrimination to be experienced in the future.

There is solid empirical evidence that expectations about rewards in the labor market influence human capital investment decisions; for example, the reward to skills held by black workers increased following the passage of the 1964 Civil Rights Act, and a rapid rise in college enrollment of blacks followed (Donohue and Heckman, 1991). But the difficulty with the argument in this context is that it presumes that black parents and children operate under mistaken expectations about the present labor market. Although it was once true that the returns to college education were lower for blacks than for whites (Becker, 1957; U.S. Civil Rights Commission, 1986), the return to college education for blacks was higher than the return for whites by the mid-1970s, and continues to be higher today. Some parallel evidence presented by Johnson and Neal (1998) shows that the returns to (coefficient on) AFQT scores for black males in an earnings equation are now as high or higher than those for whites, although they used to be lower in the pre-Civil Rights era. Given the greater return for blacks to college education and ability, it seems implausible to argue that a rational fear of lower future returns is currently discouraging black formation of skills.

Ability as it crystallizes at an early age accounts for most of the measured gap in black and white labor market outcomes. Stricter enforcement of civil rights laws is a tenuous way to improve early childhood skills and ability.[3] The weight of the evidence suggests that this ability and early motivation is most easily influenced by enriching family and preschool learning environments and by improving the quality of the early years of schooling.

The Implicit Assumptions Behind the Audit Method

The method of audit pairs operates by controlling for systematic observed differences across pairs. It does this by attempting to create two candidates for jobs or loans who are ''essentially'' the same in their paper qualifications and personal characteristics, and then comparing their outcomes in their dealings with the same firm. Averaging over the outcomes at all firms for the same audit pair produces an

[3] However, nothing I have said vindicates abolishing these laws. They have important symbolic value and they addressed and solved an important problem of blatant discrimination in the American South.

estimate of the discrimination effect. An average is often taken over audit pairs as well to report an ''overall'' estimate of discrimination. More sophisticated versions of the method will allow for some heterogeneity in treatment among firms and workers or firms and applicants.

One set of difficulties arise, however, because there are sure to be many unobserved variables. As noted by Heckman and Siegelman (1993), given the current limited state of knowledge of the determinants of productivity within firms, and given the small pools of applicants from which matched pairs are constructed that are characteristic of most audit studies, it is unlikely that all characteristics that might affect productivity will be perfectly matched. Thus, the implicit assumption in the audit pair method is that controlling for some components of productivity and sending people to the same firm will reduce the bias below what it would be if random pairs of, say, whites and blacks were compared using, for example, Census data. The implicit assumption that justifies this method is that the effect of the unobserved characteristics averages out to zero across firms for the same audit pair.

However, the mean of the differences in the unobserved components need not be zero and assuming that it is begs the problem. Nowhere in the published literature on the audit pair method will you find a demonstration that matching one subset of observable variables necessarily implies that the resulting difference in audit-adjusted treatment between blacks and whites is an unbiased measure of discrimination—or indeed, that it is even necessarily a better measure of discrimination than comparing random pairs of whites and blacks applying at the same firm or even applying to different firms. This argument is stated more formally in the Appendix. Here I present an intuitive discussion.

Consider the following example. Suppose that the market productivity of persons is determined by the sum of two productivity components. These two productivity components are distributed independently in the population so their values are not correlated with each other. Both factors affect employer assessments of employee productivity.[4] Suppose further that average productivity of the sum is the same for both whites and blacks; however, blacks are more productive on average on one component while whites are more productive on average on the other. Now consider an audit pair study that equates only the first component of productivity and equates firm effects by sending the audit pair to the same firm. Under these conditions, the audit estimator is biased toward a finding of discrimination, since in this example, only the characteristic which makes black productivity look relatively high is being used to standardize the audit pair. The condition of zero mean of unobservable productivity differences across race groups is not especially compelling and requires a priori knowledge that is typically not available.

Now consider the case in which the observed and unobserved components of

[4] They need not be perfectly observed by employers, but may only be proxied. However, it is easiest to think of both components as fully observed by the employer, but that the observing economist has less information.

productivity are dependent. In this case, making the included components as alike as possible may accentuate the differences in the unobserved components. As a result, it can increase the bias over the case where the measured components are not aligned.

A related crucial assumption behind the audit method is that it typically assumes that the outcome being studied is a linear function of the relevant variables or the outcome can be transformed to be so; that is, more skill leads in a linear way to a greater chance of employment or higher wages. However, in more general models, the decision to offer a job is a nonlinear function of the relevant characteristics. This case arises in many models of the employment decision when productivity must be above a threshold for a job offer to be made. Many models of firm hiring decisions assume that workers must possess minimum qualifications to get a job offer. A similar model is often used to explain the credit applications analyzed by Ladd; only applicants with credentials above a threshold get credit. In this case, making one set of variables more alike for an audit pair may also make the bias in estimating discrimination worse if the distributions of unobserved characteristics are not the same across race/gender groups even if the means are.

To see how this works, consider the intuition of a labor market example, patterned after the analysis of Heckman and Siegelman (1993). Again consider a case where productivity is divided into two components: one which is observable and set by the audit designer as part of deciding on the paper qualifications and personal characteristics that each member of the audit pair should possess; and the other which is unobservable to the audit study, but is at least somewhat visible to the prospective employer and acted on in hiring or credit decisions. Both are determinants of perceived (by the firm) productivity. Let us further assume that: a) the omitted and included components are statistically independent for both race groups; and b) that the mean of the unobserved productivity variable is the same for whites and blacks. Assumption b is a strong one, and it is the key identifying assumption of the audit method as currently practiced. Nothing guarantees that it will be satisfied. Under these assumptions, however, the audit methods would certainly eliminate, or at least not increase, bias in measuring racial discrimination if the measured outcomes are simple sums of productivity characteristics.

However, this conclusion only holds if the firm's treatment of the worker is linear in the attributes of the productivity level. Suppose, instead, that firms hire workers based on a common cutoff value, so that the more stringent the cutoff, the less likely is the auditor to be hired.[5] In the absence of discrimination, the cutoff for blacks will be the same as the cutoff for whites. As long as the distribution of unobserved productivity attributes is the same for each race group, hiring rates should be the same for all levels of observed characteristics set by the audit designer. Evidence of asymmetry of treatment, as manifest in the last two columns of Table

[5] It is straightforward to allow this cutoff value to vary among firms but for simplicity I assume it does not. If these differences in cutoff values are random across firms and pairs, the analysis in the text goes through without any essential modification.

1, would be evidence that the firm cutoff rates are different among race groups, and thus that a different standard is applied to whites than to blacks.

Suppose, however, that whites are more heterogeneous in their observed productivity characteristics; as a result, the distribution of their unobserved productivity characteristics have different tail areas. In this case, the ability of the audit pair method to detect discrimination will depend on the level at which the observed level of productivity was standardized in the audit design.

Here is the intuitive argument. Say that blacks and whites face the same productivity threshold for hiring, and have the same mean characteristics, but that whites are more heterogeneous in their productivity characteristics. For simplicity, assume that these characteristics are symmetrically distributed. Imagine that the audit team members are applying for a job where their observed characteristics, as set by the audit design, are such that they are highly qualified for the job. Then the prospective employer, who has insight into the characteristics unobserved by the designer of the audit study, will tend to hire more blacks than whites. In effect, the longer left-hand tail of unobserved characteristics for whites works against them.[6] Precisely the opposite bias appears if black unobservables are more dispersed than white unobservables. Now consider applying for a job where the observed characteristics, as set by the audit design, are such that the auditors have only a moderate chance of qualifying for the job. Thus most of the applicants of both race groups are below the threshold. Then the prospective employer, who has insight into the characteristics unobserved by the designer of the audit study, will tend to hire more whites than blacks if white traits are more dispersed. In effect, the longer right-hand tail of the distribution of unobserved characteristics for whites works for them.[7]

Another way to state this point is to think of pairing up black and white high jumpers to see if they can clear a bar set at a certain height. There is no discrimination, in the sense that they both use the same equipment and have the bar set at the same level. Suppose now that the chance of a jumper (of any race) clearing the bar depends on two additive factors: the person's height and their jumping technique. We can pair up black and white jumpers so that they have identical heights, but we can't directly observe their technique. Let us make the generous assumption, implicit in the entire audit literature, that the mean jumping technique is equal for the two groups. Then, if the variance of technique is also the same for white and black high-jumpers, we would find that the two racial groups are equally likely to clear the bar. On the other hand, if the variance differs, then whether the black or white pair is more likely to clear the bar will depend on how the bar is set, relative to their common height, and which racial group has a higher variance in jumping technique. If the bar is set at a low level so that most people of the given height are likely to clear the bar, then the group with the lower variance will be

[6] If the black distribution of unobservables is more heterogeneous, whites are favored. This interpretation is consistent with the evidence in Table 1.

[7] This case seems to rationalize the evidence in Table 1.

more likely to clear the bar. If the bar is set at a very high level relative to the given height, then the group with a higher variance in jumping technique will be more likely to clear the bar. A limitation of the audit method is readily apparent from this analogy: there is no discrimination, yet the two groups have different probabilities of clearing the bar.[8] And if there is discrimination—that is, the bar is being set higher for blacks—the differential dispersion in the unobserved component could still cause the minority group to clear the bar more often. The method could fail to detect discrimination when it does exist.

Thus, depending on the distribution of unobserved characteristics for each race group and the audit standardization level, the audit method can show reverse discrimination, or equal treatment, or discrimination, even though blacks and whites in this example are subject to the same cutoff and face no discrimination. The apparent bias depends on whether the level of qualifications set by the audit designer makes it more or less likely that the applicant will receive the job, and the distribution of variables that are unobservable to the audit design. The apparent disparity favoring Washington whites in Table 1 may be a consequence of differences in unobserved characteristics between blacks and whites when there is no discrimination.

Even more disturbing, suppose that there is discrimination against blacks, so the productivity cutoff used by firms is higher for blacks than whites. Depending on the audit designer's choice of what level of qualifications are given to the auditors, the audit study can find no discrimination at all. However, whether the qualifications make it relatively likely or unlikely to get the job is a fact rarely reported in audit studies. This argument is made more precise in the Appendix where two examples are given, where discrimination or reverse discrimination is detected when none exists, and discrimination is disguised when it exists.

Making audit pairs as alike as possible may seem an obviously useful step, but it can greatly bias the inference about average discrimination or discrimination at the margin. Intuitively, by taking out the common components that are most easily measured, differences in hiring rates as monitored by audits arise from the idiosyncratic factors, and not the main factors, that drive actual labor markets. These examples highlight the fragility of the audit method to untested and unverifiable assumptions about the distributions of unobservables. Similar points arise in more general nonlinear models that characterize other employment decision rules.

The Becker Model

The papers in this symposium make the erroneous claim that in Becker's (1957) model, market discrimination disappears in the long run. It need not. Entrepreneurs

[8] I owe this analogy to Alan Krueger. This analogy also shows how artificial the audit studies are because one would expect to find athletes choosing their sports based on their chances of success, as in the purposive search in the labor market discussed earlier.

can consume their income in any way they see fit. If a bigoted employer prefers whites, the employer can indulge that taste as long as income is received from entrepreneurial activity, just as a person who favors an exotic ice cream can indulge that preference by being willing to pay the price. Only if the supply of entrepreneurship is perfectly elastic in the long run at a zero price, so entrepreneurs have no income to spend to indulge their tastes, or if there are enough nonprejudiced employers to hire all blacks, will discrimination disappear from Becker's model.

However, even if the common misinterpretation of Becker's model is accepted, it is far from clear that the prediction of no or little discrimination in the U.S. labor market in the long run is false. The substantial decline over the past 50 years in wage differentials between blacks and whites may well be a manifestation of the dynamics of the Becker model. It may take decades for the effects of past discrimination in employment and schooling as it affects current endowments of workers to fade out of the labor market. But the evidence from the current U.S. labor market is that discrimination by employers alone does *not* generate large economic disparities between blacks and whites.

Appendix

Implicit Identifying Assumptions In The Audit Method

Define the productivity of a person of race $r \in \{1, 0\}$, at firm f, with characteristics $\underset{\sim}{X} = (X_1, X_2)$ as $P(\underset{\sim}{X}, r, f)$. $r = 1$ corresponds to black; $r = 0$ corresponds to white. Assume that race does not affect productivity so we may write $P = P(\underset{\sim}{X}, f)$. The treatment at the firm f for a person of race r and productivity P is $T(P(\underset{\sim}{X}, f), r)$. Racial discrimination exists at firm f if

$$T(P(\underset{\sim}{X}, f), r = 1) \neq \mathrm{T}(P(\underset{\sim}{X}, f), r = 0).$$

As noted in the text, audit methods monitor discrimination at randomly selected firms within the universe designated for sampling, not the firms where blacks are employed.

The most favorable case for auditing assumes that T (or some transformation of it) is linear in f and X. Assume for simplicity that $P = X_1 + X_2 + f$ and $T(P, r) = P + \gamma r$. When $\gamma < 0$ there is discrimination against blacks. γ may vary among firms as in Heckman and Siegelman (1993).[9] For simplicity suppose that all firms are alike. Audit methods pair racially dissimilar workers in the following way: they match

[9] For simplicity, assume that γ is the same across all firms. Alternatively, assume that it is distributed independently of $\underset{\sim}{X}$ and f.

some components of $\underset{\sim}{X}$ and they sample the same firms. Let P_1^* be the standardized productivity for the black member of the pair; P_0^* is the standardized productivity for the white member. If $P_0^* = P_1^*$,

$$T(P_1^*, 1) - T(P_0^*, 0) = \gamma.$$

When averaged over firms, the average treatment estimates the average γ.

Suppose that standardization is incomplete. We can align the first coordinate of X at $\{X_1 = X_1^*\}$ but not the second coordinate, X_2, which is unobserved by the auditor but acted on by the firm. $P_1^* = X_1^* + X_2^1$ where X_2^1 is the value of X_2 for the $r = 1$ member and $P_0^* = X_1^* + X_2^1$. In this case

$$T(P_1^*, 1) - T(P_0^*, 0) = X_2^1 - X_2^0 + \gamma.$$

For averages over pairs to estimate γ without bias, it must be assumed that $E(X_2^1) = E(X_2^0)$; i.e., that the mean of the unobserved productivity traits is the same. This is the crucial identifying assumption in the conventional audit method. Suppose that this is true so $E(X_2^1) = E(X_2^0) = \mu$. Then the pair matching as in the audit method does not increase bias and in general reduces it over comparisons of two X_1-identical persons at two randomly selected firms. Under these conditions, bias is lower than if two randomly chosen auditors are selected at the same firm if $E(X_1^1) \neq E(X_1^0)$.

However, the decision rule to offer a job or extend credit often depends on whether or not the perceived productivity P exceeds a threshold c:

$$T = 1 \text{ if } P \geq c$$

$$T = 0 \text{ otherwise.}$$

In this case, the audit pair method will still produce bias even when it does not when T is linear in $\underset{\sim}{X}$ and f, unless the *distributions* of the omitted characteristics are identical in the two race groups. Suppose that $P = X_1 + X_2$. X_2 is uncontrolled. Then assuming no discrimination ($\gamma = 0$)

$$\begin{aligned} T(P_1^*, 1) &= 1 \text{ if } X_1^* + X_2^1 + f \geq c \\ &= 0 \text{ otherwise} \\ T(P_0^*, 0) &= 1 \text{ if } X_1^* + X_2^0 + f \geq c \\ &= 0 \text{ otherwise.} \end{aligned}$$

Even if the distributions of f are identical across pairs, and f is independent of X, unless the *distributions* of X_2^1 and X_2^0 are identical, $\Pr(T(P_1^*, 1) = 1) \neq$

Figure 1

Relative Hiring Rate as a Function of the Level of Standardization. Blacks Have More Dispersion. Threshold Hiring Rule: No Discrimination Against Blacks Normally Distributed Unobservables

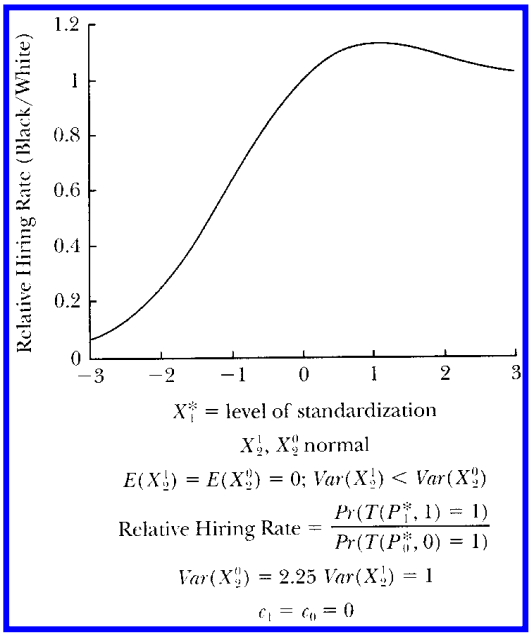
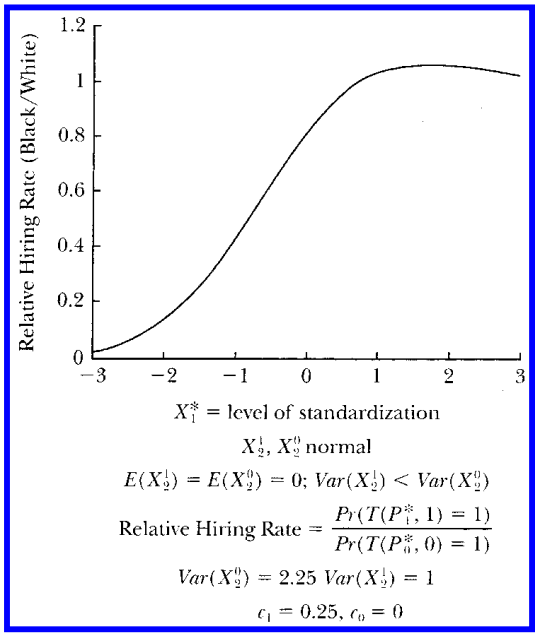


Figure 2

Relative Hiring Rate as a Function of the Level of Standardization. Blacks Held to Higher Standard; Blacks Have More Dispersion. Threshold Hiring Rule: No Discrimination against Blacks Normally Distributed Unobservables



$\Pr(T(P_0^*, 0) = 1)$ for most values of the standardization level X_1^*. The right tail area of the distribution governs the behavior of these probabilities. This implies that even if blacks and whites face the same cutoff value, and in this sense are treated without discrimination in the labor market, even if the means of the distributions of unobservables are the same across race group, if the distributions of the unobservables are different, their probabilities of being hired will differ and will depend on the level of standardization used in the audit study—something that is rarely reported. The pattern of racial disparity in Table 1 may simply be a consequence of the choice of the level of standardization in those audits, and not discrimination.

Worse yet, suppose that the cutoff $c = c_1$ for blacks is larger than the cutoff $c = c_0$ for whites so that blacks are held to a higher standard. Then depending on the right tail area of X_2^1 and X_2^0, the values of c_1 and c_0, and the level of standardization X_1^*,

$$\Pr(T(P_1^*, 1) = 1) \gtreqless P(T(P_0^*, 0) = 1).$$

In general, only if the *distributions* of X_2^1 and X_2^0 are the same for each race group,

will the evidence reported in Table 1 be informative on the level of discrimination in the universe of sampled firms.

Figures 1 and 2 illustrate these two cases for X_2^1 and X_2^0 normally distributed (and independent of each other) where X_1^* is the level of audit standardization and firms are standardized to have $f = 0$.[10] In Figure 1 there is no discrimination in the market. Yet the black hire rate falls short of the white rate if the standardization rate is $X_1^* < 0$, and the lower the value of X_1^*, the greater the shortfall. In Figure 2, which is constructed for a hypothetical economy where there is discrimination against blacks, for high standardization rates, audits would appear to reveal discrimination *in favor* of blacks when in fact blacks are being held to a higher standard. The evidence in Table 1 is intrinsically ambiguous about the extent of discrimination in the market. For further discussion, see Heckman and Siegelman (1993).

■ *This research is supported by a grant from the American Bar Foundation. I thank Mark Killingsworth for references and Alan Krueger, Derek Neal and Timothy Taylor for valuable comments.*

[10] Allowing f to vary but assuming it is normal mean zero and variance σ_f^2 does not change the qualitative character of these calculations assuming that f is distributed independently of the characteristics.

References

Abowd, John, and Mark Killingsworth, ''Sex, Discrimination, Atrophy, and the Male-Female Wage Differential,'' *Industrial And Labor Relations Review*, Fall 1983, *22*:3, 387–402.

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

Cameron, Stephen, and James Heckman, ''The Nonequivalence of High School Equivalents,'' *Journal of Labor Economics*, 1993, *11*:1, pt1, 1–47.

Cawley, John, James Heckman, and Edward Vytlacil, ''Cognitive Ability and the Rising Return to Education,'' NBER working paper 6388, January 1998.

Donohue, John, and James Heckman, ''Continuous vs. Episodic Change: The Impact of Affirmative Action and Civil Rights Policy on The Economic Status of Blacks,'' *Journal of Economic Literature*, December 1991, *29*:4, 1603–43.

Heckman, James, ''The Central Role of the South in Accounting For The Economic Progress of Black Americans,'' Papers and Proceedings Of The American Economic Association, May 1990.

Heckman, James, ''Lessons From the Bell Curve,'' *Journal of Political Economy*, 1995, *103*:5, 1091–1120.

Heckman, James, and Peter Siegelman, ''The Urban Institute Audit Studies: Their Methods and Findings.'' In M. Fix and R. Struyk, eds. *Clear and Convincing Evidence: Measurement of Discrimination in America*. Urban Institute, Fall 1993.

Heckman, James, Lance Lochner, Christopher Taber, and Jeffrey Smith, ''The Effects of Government Policy on Human Capital Investment and Wage Inequality,'' *Chicago Policy Review*, Spring 1997, *1*:2, 1–40.

Johnson, William R., and Derek Neal, ''Basic Skills and the Black-White Earnings Gaps.'' In Jencks, Christopher, and Meredith Phillips, eds. *The Black-White Test Score Gap*. Washington, D.C. Brookings, 1998.

Neal, Derek, and William Johnson, ''The Role of Premarket Factors in Black-White Wage Differences,'' *Journal of Political Economy*, 1996, *104*:5, 869–95.

Rodgers III, William, and William Spriggs, ''What Does AFQT Really Measure: Race, Wages, Schooling and the AFQT Score,'' *The Review of Black Political Economy*, Spring 1996, *24*:4, 13–46.

Rodgers III, William, William E. Spriggs, and Elizabeth Waaler, ''The Role of Premarked Factors in Black-White Differences: Comment,'' Unpublished Manuscript, College of William and Mary, May 25, 1997.

Seelye, Katherine, ''Employment Panel To Send People Undercover to Detect Bias in Hiring,'' *New York Times*, Sunday, December 7, 1997, p. 22.

U.S. Commission on Civil Rights, *The Economic Progress of Black Men in America*, Clearinghouse Publication 91, 1986.

This article has been cited by:

1. David Neumark. 2018. Experimental Research on Labor Market Discrimination. *Journal of Economic Literature* **56**:3, 799-866. [Abstract] [View PDF article] [PDF with links]
2. David Neumark, Ian Burn, Patrick Button. 2016. Experimental Age Discrimination Evidence and the Heckman Critique. *American Economic Review* **106**:5, 303-308. [Abstract] [View PDF article] [PDF with links]
3. Kevin Lang,, Jee-Yeon K. Lehmann. 2012. Racial Discrimination in the Labor Market: Theory and Empirics. *Journal of Economic Literature* **50**:4, 959-1006. [Abstract] [View PDF article] [PDF with links]
4. Eric Zitzewitz. 2012. Forensic Economics. *Journal of Economic Literature* **50**:3, 731-769. [Abstract] [View PDF article] [PDF with links]
5. Kevin Lang,, Michael Manove. 2011. Education and Labor Market Discrimination. *American Economic Review* **101**:4, 1467-1496. [Abstract] [View PDF article] [PDF with links]
6. Markus M. Mobius, Tanya S. Rosenblat. 2006. Why Beauty Matters. *American Economic Review* **96**:1, 222-235. [Abstract] [View PDF article] [PDF with links]
7. Steven N. Durlauf. 2005. Racial Profiling as a Public Policy Question: Efficiency, Equity, and Ambiguity. *American Economic Review* **95**:2, 132-136. [Citation] [View PDF article] [PDF with links]
8. Glenn W. Harrison, John A. List. 2004. Field Experiments. *Journal of Economic Literature* **42**:4, 1009-1055. [Abstract] [View PDF article] [PDF with links]
9. Marianne Bertrand, 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. [Abstract] [View PDF article] [PDF with links]
10. Uri Gneezy, Aldo Rustichini. 2004. Gender and Competition at a Young Age. *American Economic Review* **94**:2, 377-381. [Citation] [View PDF article] [PDF with links]
11. Harry Holzer,, David Neumark. 2000. Assessing Affirmative Action. *Journal of Economic Literature* **38**:3, 483-568. [Abstract] [View PDF article] [PDF with links]
12. Richard A. Posner,. 1999. The Law and Economics of the Economic Expert Witness. *Journal of Economic Perspectives* **13**:2, 91-100. [Citation] [View PDF article] [PDF with links]

E-lit-014: Detecting Discrimination in Audit and Correspondence Studies

David Neumark | Journal of Human Resources 47(4): 1128-1157 (NBER Working Paper 16448) | 2012

https://www.nber.org/system/files/working_papers/w16448/w16448.pdf

About this source

Neumark accepts the strongest form of the Heckman-Siegelman critique: even a perfect correspondence study, with identical observed and unobserved group means, can generate spurious evidence of discrimination in either direction if the variance of unobserved productivity differs by group, because hiring is a threshold decision and the sign of the bias flips with the level at which applicant quality is standardized. He then shows the problem is solvable when a study varies applicant quality in ways that move callbacks: the ratio of the effect of qualifications on callbacks for one group to the other identifies the relative variance of the unobservable, which in turn identifies the race coefficient, estimable as a heteroskedastic probit. Bertrand and Mullainathan's design happens to satisfy this requirement, because it uses low- and high-quality resumes, so their data can be re-analyzed. The raw probit gap of 3 to 3.3 percentage points (about 33 percent of the 9.65 percent white callback rate) falls to an overall marginal effect of 2.4 to 2.5 points once the variance is allowed to differ by race — but the decomposition matters more than the total: the effect of race through the level of the latent productivity variable, which is the part that counts as discrimination and generalizes outside the experiment, is LARGER in absolute value (-0.054 to -0.086), while a positive and statistically insignificant effect through the variance (0.028 to 0.062) partly offsets it. The estimated standard deviation of the unobservable is 1.15 to 1.41 times larger for blacks than for whites, above one but never significantly so, and the overidentifying restrictions implied by the identifying assumption are never rejected (p-values 0.17 to 0.62, and 0.68 to 0.92 in the restricted version). Monte Carlo simulations confirm that a plain probit can manufacture marginal effects of plus or minus 0.1 with zero true discrimination, that the heteroskedastic probit removes that bias, and that when the identifying assumption is violated the remaining bias is multiplicative — it can distort the magnitude but not the sign. For the section's hypothesis, this is the pivotal paper: the critique is real, and the correction it motivates turns the flagship modern correspondence study's finding UP rather than overturning it, precisely because that study randomized applicant quality within the design.

Bound as the original PDF. Retrieved 2026-09-10. sha256 60cf7b221a1d1dae...

Bound into an Economic Briefing · by Claude, may contain mistakes

Full document begins on the next page.

NBER WORKING PAPER SERIES

DETECTING DISCRIMINATION IN AUDIT AND CORRESPONDENCE STUDIES

David Neumark

Working Paper 16448
http://www.nber.org/papers/w16448

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2010

David Neumark is Professor of Economics and Director of the Center for Economics & Public Policy at UCI, a research associate of the NBER, and a research fellow at IZA. He is grateful to the UCI Academic Senate Council on Research, Computing, and Libraries for research support, to Scott Barkowski, Marianne Bitler, Richard Blundell, Thomas Cornelißen, Ying-Ying Dong, Judith Hellerstein, James Heckman, an anonymous referee, and seminar participants at Baylor University, Cal State Fullerton, Cornell University, the Federal Reserve Bank of San Francisco, Hebrew University, the Melbourne Institute, the University of Oklahoma, Tel Aviv University, the University of Sydney, and the All-California Labor Conference for helpful comments, and to Scott Barkowski, Andrew Chang, Jennifer Graves, and Smith Williams for research assistance. He also thanks Marianne Bertrand and Sendhil Mullainathan for supplying their data, available from the AEA website at http://www.aeaweb.org/articles.php?doi=10.1257/0002828042002561.

The simulation data and the computer code used in this paper can be obtained beginning six months after publication through three years hence from David Neumark, Department of Economics, 3151 Social Science Plaza, UCI, Irvine, CA, 92697, dneumark@uci.edu. The views expressed herein are those of the author 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.

© 2010 by David Neumark. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Detecting Discrimination in Audit and Correspondence Studies
David Neumark
NBER Working Paper No. 16448
October 2010, Revised September 2012
JEL No. C93,J7

ABSTRACT

Audit studies testing for discrimination have been criticized because applicants from different groups may not appear identical to employers. Correspondence studies address this criticism by using fictitious paper applicants whose qualifications can be made identical across groups. However, Heckman and Siegelman (1993) show that group differences in the variance of unobservable determinants of productivity can still generate spurious evidence of discrimination in either direction. This paper shows how to recover an unbiased estimate of discrimination when the correspondence study includes variation in applicant characteristics that affect hiring. The method is applied to actual data and assessed using Monte Carlo methods.

David Neumark
Department of Economics
University of California at Irvine
3151 Social Science Plaza
Irvine, CA 92697
and NBER
dneumark@uci.edu

I. Introduction

In audit or correspondence studies, fictitious individuals who are identical except for race, sex, or ethnicity apply for jobs. Group differences in outcomes – for example, blacks getting fewer job offers than whites – are interpreted as reflecting discrimination. Across a wide array of countries and demographic groups, audit or correspondence studies find evidence consistent with discrimination, including discrimination against blacks, Hispanics, and women in the United States (Mincy 1993; Neumark 1996; Bertrand and Mullainathan [BM] 2004), Moroccans in Belgium and the Netherlands (Smeeters and Nayer 1998; Bovenkerk, Gras, and Ramsoedh 1995), and lower castes in India (Banerjee et al. 2008). These "field experiments" are widely viewed as providing the most convincing evidence on discrimination (Pager 2007; Riach and Rich 2002), and U.S. courts allow organizations that conduct audit or correspondence studies to file claims of discrimination based on the evidence they collect (U.S. Equal Employment Opportunity Commission 1996). Yet audit or correspondence studies have been sharply criticized by Heckman and Siegelman [HS] (HS 1993; Heckman 1998). Perhaps the most damaging criticism is that when the variance of unobserved productivity differs across groups – as in the standard statistical discrimination model (Aigner and Cain 1977) – audit or correspondence studies can generate spurious evidence of discrimination in either direction or of its absence; equivalently, discrimination is unidentified in these studies. Although this critique has been ignored in the literature, it clearly casts serious doubt on the validity of the evidence from these studies. This paper addresses the unobserved variance critique of audit and correspondence studies, proposing a method of collecting and analyzing data from these studies that can correctly identify discrimination.

The HS criticism that has received the most attention is that audit studies – which use "live" job applicants – fail to ensure that applicants from different groups appear identical to

employers. Many of these criticisms can be countered by using correspondence studies, which use fictitious applicants on paper, or more recently the internet, whose qualifications can be made identical across groups. However, the Heckman and Siegelman (1993) critique applies even in the ideal case in which both observed and unobserved group averages are identical.

This paper develops and implements a method of using data from audit or correspondence studies that accomplishes two goals. First, it provides a statistical test of whether HS's unobservable variance critique applies to the data from a particular study. Second, and more important, it develops a statistical estimation procedure that identifies the effect of discrimination. The method requires the study to have variation in applicant characteristics that affect hiring. It is a simple matter to collect the requisite data in future correspondence studies, and as an illustration the method is implemented using data from a correspondence study (BM 2004) that has the requisite data.

The method rests on three types of assumptions. First, it is based on an assumed binary threshold model of hiring that asks whether the perceived productivity of a worker exceeds a standard. Second, it imposes a parametric assumption about the distribution of unobservables that is necessary for identification in this case. The model and parametric assumption parallel exactly the setting used by HS to interpret data from these types of studies; but there may of course be other contexts where this method is useful. Finally, to solve the identification problem highlighted by HS, it relies on an additional identifying assumption that some applicant characteristics affect the perceived productivity of workers, and hence hiring, and that the effects of these characteristics on perceived productivity do not vary with group membership (for example, race). This identifying assumption has testable implications in the form of overidentifying restrictions. The estimation procedure is assessed via Monte Carlo simulations.

II. Background on Audit and Correspondence Studies

Earlier research on labor market discrimination focused on individual-level employment or earnings regressions, with discrimination estimated from the race, sex, or ethnic differential that remains unexplained after including many proxies for productivity. These analyses suffer from the obvious criticism that the proxies do not adequately capture group differences in productivity, in which case the "unexplained" differences cannot be interpreted as discrimination.

Audit or correspondence studies are a response to this inherent weakness of the regression approach to discrimination. These studies are based on comparisons of outcomes (usually job interviews or job offers) for matched job applicants differing by race, sex, or ethnicity (see, for example, Turner, Fix, and Struyk 1991; Neumark 1996; BM 2004). Audit or correspondence studies directly address the problem of missing data on productivity. Rather than trying to control for variables that might be associated with productivity differences between groups, these studies instead create an artificial pool of job applicants, among whom there are intended to be no average differences by group. By using either applicants coached to act alike, with identical-quality resumes (an audit study), or simply applicants on paper who have equal qualifications (a correspondence study), the method is largely immune to criticisms of failure to control for important differences between, for example, black and white job applicants. As a consequence, this strategy has come to be widely used in testing for discrimination in labor markets (as well as housing markets). Thorough reviews are contained in Fix and Struyk (1993), Riach and Rich (2002), and Pager (2007).

Despite the widely-held view that audit or correspondence studies are the best way to test for labor market discrimination, critiques of these studies challenge their conclusions (HS 1993; Heckman 1998). Many of these criticisms have been acknowledged by researchers as potentially

valid, and subsequent research has adapted. For example, HS noted that in the prominent audit studies carried out by the Urban Institute (for example, Mincy 1993), white and minority testers were told, during their training, about "the pervasive problem of discrimination in the United States," raising the possibility that testers subconsciously took actions in their job interviews that led to the "expected" result (HS 1993). A constructive response to this criticism has been the move to correspondence studies, which focus on applications on paper and whether they result in job interviews, thus cutting out the influence of the individual job applicants used in the test.

However, a fundamental critique of audit or correspondence studies has not been addressed by researchers. In particular, HS consider what most researchers view as the ideal conditions for an audit or correspondence study – when not only are the observable average differences between groups eliminated, but in addition the observable characteristics used in the applications are sufficiently rich that it is reasonable to assume that potential employers believe there are no average differences in *unobservable* characteristics across groups. HS show that, even in this case, audit or correspondence studies can generate evidence of discrimination (in either direction) when there is none, and can also mask evidence of discrimination when it in fact exists. Given the pervasive use of audit and correspondence studies, the failure of any research to address this critique is a significant gap in the social science and legal literature.

III. The Heckman-Siegelman Critique

I first set up the analytical framework that parallels the one used in HS's critique of evidence from audit and correspondence studies. This framework is used to illustrate the problem of identifying discrimination when there are group differences in the variance of unobservables. Then, in the following section, I show how discrimination can be identified in this setting.

Suppose that productivity depends on two individual characteristics, $X' = (X^I, X^{II})$. Let R

be a dummy for race, with R = 1 for minorities and 0 for non-minorities (which I will refer to as "black" and "white" for short). Allow productivity also to depend on a firm-level characteristic F, so that productivity is P(X',F). Let the treatment of a worker depending on P and possibly R (if there is discrimination) be denoted T(P(X',F),R). For now, think of this treatment as continuous, even though that is not the usual outcome for an audit or correspondence study; suppose the treatment is, for example, the wage offered, set equal to a worker's productivity minus a possible discriminatory penalty for blacks as in Becker's (1971) employer taste discrimination model.

Define discrimination as

(1) $T(P(X',F)|R = 1) \neq T(P(X',F)|R = 0)$.

Assume that P(.,.) and T(P(.,.)) are additive, so

(2) $P(X',F) = \beta_I' X^I + X^{II} + F$

(3) $T(P(X',F),R) = P + \gamma' R$.[1]

Thus, discrimination against blacks implies that $\gamma' < 0$, so that blacks are paid less than equally-productive whites at the same firm.

In an audit or correspondence study, two testers (or applications) or multiple pairs of testers (one with R = 1 and one with R = 0 in each pair) are sent to firms to apply for jobs. The researcher attempts to standardize their productivity based on observable productivity-related characteristics. Denote expected productivity for blacks and whites, based on the productivity-related characteristics that the firm observes, as P_B^* and P_W^*; these are not necessarily based on both X^I and X^{II}, as we may want to treat X^{II} as unobserved by firms. The goal of the audit or correspondence study design is to set $P_B^* = P_W^*$. Given these observables, the outcome T is observed for each tester. So based on Equation 3, each test – thought of as the outcome of applications to a firm by one black and one white tester – yields an observation

(4) $T(P_B^*,1) - T(P_W^*,0) = P_B^* + \gamma' - P_W^*$.

If $P_B^* = P_W^*$, then averaging across tests yields an estimate of γ'. In this case, we can also estimate the mean difference between the outcome T for blacks and white by a regression of T on a constant and the race dummy, or

(5) $T(R) = \alpha' + \gamma' R_i + \varepsilon_i$

where the first argument of T(.,.) is suppressed because it is assumed the same for all applicants.[2]

Now consider explicitly the two components of productivity, X^I and X^{II}. Suppose the audit or correspondence study controls only X^I in the resumes or interviews.[3] Denote by X_B^j and X_W^j the values of X^I and X^{II} for blacks and whites, j = I, II. Suppose that the audit or correspondence study, as is usually done, sets $X_B^I = X_W^I$; the level at which they are set is later denoted as X^{I*}, the level at which X^I is "standardized" across applicants. Then for the test resulting from the application of a pair of black and white testers to a firm, P_B^* and P_W^* are

(6) $P_B^* = \beta_I' X_B^I + E(X_B^{II}) + F$

(7) $P_W^* = \beta_I' X_W^I + E(X_W^{II}) + F$.

(For simplicity, I suppress the conditioning on X_B^I or X_W^I. Both P_k^* and $E(X_k^{II})$, k = B, W, can be interpreted as conditional on X_k^I.)

In this case, each individual test provides an observation equal to

(8) $T(P_B^*,1) - T(P_W^*,0) = P_B^* + \gamma' - P_W^* = \beta_I' X_B^I + E(X_B^{II}) + \gamma' - (\beta_I' X_W^I + E(X_W^{II}))$

$= \gamma' + E(X_B^{II}) - E(X_W^{II})$.

Clearly observations from a sample of such tests identify γ' only if $E(X_B^{II}) = E(X_W^{II})$. Thus, a key assumption in an audit or correspondence study is that all productivity-related factors not controlled for in the test have the same mean for blacks and whites. Heckman (1998) and HS (1993), in critiques unrelated to the focus of this paper (group differences in the variance

of unobservables), offer a detailed discussion of the reasons why this assumption might be violated in audit studies. Specifically, despite researchers' best efforts to standardize applicants, differences remain that may be observed by employers. And HS further show that even when these differences are weakly related to productivity, they can lead to large biases, precisely *because* the applicants are standardized on other productivity-related characteristics. Moreover, the design of an audit study allows for experimenter effects that can – through information conveyed to employers in the job interview – generate differences between $E(X_B^{II})$ and $E(X_W^{II})$.

Correspondence studies are a response to these criticisms of audit studies (see, for example, BM, p. 994). In contrast to audit studies, they do not entail face-to-face interviews that might convey mean differences on uncontrolled variables between blacks and whites, and hence the researcher can eliminate differences observed by employers that are not controlled in the study. In addition, a correspondence study, by its very nature, avoids potential experimenter effects.

Even in a correspondence study, though, differences in employer estimates of mean unobserved characteristics for blacks and whites can affect the results, as in Equation 8. The difference, in this case, is one of interpretation, and even when this possibility arises, correspondence studies still have an important advantage. Regarding interpretation, because employers are not allowed to make assumptions about race or sex differences in characteristics not observed in the job application or interview process (U.S. Equal Employment Opportunity Commission, n.d.), any role of assumed mean differences in characteristics in affecting the outcomes from a correspondence study can be interpreted as statistical discrimination. Consequently, we can interpret the estimate of the expression in Equation 8 from a correspondence study as capturing the combined effects of taste discrimination (γ') and statistical discrimination ($E(X_B^{II}) - E(X_W^{II})$). But in a correspondence study, the estimate of the combined

effects of the two types of discrimination is still more reliable than in an audit study, because of the absence of experimenter effects.

In other words, correspondence studies succeed, where audit studies may fail, in providing unbiased estimates of what the law recognizes as discrimination. However, they are not necessarily better at isolating taste discrimination. Nonetheless, when a correspondence study includes a rich set of applicant characteristics, it becomes less likely that statistical discrimination plays much of a role in group differences in outcomes.[4]

Turning to the main focus of this paper, HS show that a more troubling result emerges in audit or correspondence studies because the relevant treatment is not linear in productivity as it might be for a wage offer, but instead is non-linear. That is, we think that in the hiring process firms evaluate a job applicant's productivity relative to a standard, and offer the applicant a job (or an interview) if the standard is met. In this case, HS show that, even when there are equal group averages of both observed and unobserved variables, an audit or correspondence study can generate biased estimates, with spurious evidence of discrimination in either direction, or of its absence – or, in other words, discrimination is unidentified. Because this critique applies even to correspondence studies, which meet higher standards of validity, the remainder of the discussion refers exclusively to correspondence studies.

The intuitive basis of the HS critique is as follows. Consider the simplest case in which $E(X_B^j) = E(X_W^j)$, j = I, II, and the only difference between blacks and whites is that the *variance* of unobserved productivity is higher for whites than for blacks, for example. The correspondence study controls for one productivity-related characteristic, X^I, and standardizes on a quite low value of X^I (that is, the study makes the two groups equal on characteristic X^I, but at a low value X^{I*}). The correspondence study does not convey any information on a second, unobservable productivity-related characteristic, X^{II}. Because an employer will offer a job

interview only if it perceives or expects the sum $\beta_I'X^I + X^{II}$ to be sufficiently high, when X^{I*} is set at a low level the employer has to believe that X^{II} is high (or likely to be high) in order to offer an interview. Even though the employer does not observe X^{II}, if the employer knows that the variance of X^{II} is higher for whites, the employer correctly concludes that whites are more likely than blacks to have a sufficiently high sum of $\beta_I'X^I + X^{II}$, by virtue of the simple fact that fewer blacks have very high values of X^{II}. Employers will therefore be less likely to offer jobs to blacks than to whites, *even though the observed average of X^I is the same for blacks and whites, as is the unobserved average of X^{II}*. The opposite holds if the standardization is at a high value of X^I; in the latter case the employer only needs to *avoid* very low values of X^{II}, which will be more common for the higher-variance whites.

The idea that the variances of unobservables differ across groups has a long tradition in research on discrimination, stemming from early models of statistical discrimination. For example, Aigner and Cain (1977) discuss these models and suggest that a higher variance of unobservables for blacks compared to whites is plausible, and Lundberg and Startz (1983) study how such an assumption can lead to an equilibrium with lower investment in human capital by blacks. On the other hand, Neumark (1999) finds no evidence that employers have better labor market information about whites than blacks, and if anything the estimates – although imprecise – indicate the opposite.

To see the bias formally, suppose that a job offer or interview is given if a worker's perceived productivity exceeds a certain threshold c'. As before, suppose that P is determined as a linear sum of X^I, X^{II}, and F (Equation 2), with X^{II} (and F) statistically independent of X^I.[5] The hiring rules for blacks and whites (with the possibility of discrimination) are

(9) $T(P(X^{I*},X_B{}^{II})|R = 1) = 1$ if $\beta_I'X^{I*} + X_B{}^{II} + \gamma' + F > c'$

(9') $T(P(X^{I*},X_W{}^{II})|R = 0) = 1$ if $\beta_I'X^{I*} + X_W{}^{II} + F > c'$.

Discrimination may lead employers to "discount" the productivity of a black worker, as captured in γ'.

To get to an econometric specification, assume that the unobservables X_B^{II} and X_W^{II} are normally distributed, with equal means (set to zero, without loss of generality), and standard deviations σ_B^{II} and σ_W^{II}. Finally, as long as the firm-specific productivity shifters F are normally distributed and independent of X^{II}, and because they have the same distribution for blacks and whites, we can ignore them and focus solely on the unobserved variation in X^{II} (effectively redefining the random variable in what follows as $X^{II} + F$). Under these assumptions, the probabilities that blacks and whites get hired are

(10) $\Pr[T(P(X^{I*},X_B^{II})|R = 1) = 1] = 1 - \Phi[(c' - \beta_I'X^{I*} - \gamma')/\sigma_B^{II}] = \Phi[(\beta_I'X^{I*} + \gamma' - c')/\sigma_B^{II}]$

(10') $\Pr[T(P(X^{I*},X_W^{II}) |R = 0) = 1] = 1 - \Phi[(c' - \beta_I'X^{I*})/\sigma_W^{II}] = \Phi[(\beta_I'X^{I*} - c')/\sigma_W^{II}]$,

where Φ denotes the standard normal distribution function.

There is a potential difference between audit and correspondence studies in terms of how we might think about what is observable to the econometrician and the firm. In an audit study, the simplest characterization might by that both the firm and the econometrician observe X_B^I and X_W^I, but the firm also observes X_B^{II} and X_W^{II}. In a correspondence study, however, both the firm and the econometrician observe only X_B^I and X_W^I, with no variation in X_B^{II} and X_W^{II} observed by employers. However, the distinction between audit and correspondence studies is not quite this sharp, because even though an employer interviews an applicant in an audit study, some (perhaps many) determinants of productivity remain unobserved when a job offer is made.

This issue is relevant to thinking about how we arrive at a statistical model of hiring, such as the probit specifications in Equations 10 and 10'. In an audit study, variables unobserved by the econometrician but observed by the firm can generate variation in hiring. In a correspondence study, however, if firms observe only X_B^I and X_W^I, then the decision about who

to hire should be deterministic. Given X_B^I and X_W^I (assumed equal), the employer hires the higher variance group if the level of standardization is low, and vice versa, so all firms evaluating identical job applicants should make the same decisions about hiring whites and blacks. One way to introduce unobservables that generate random variation, with variances that differ by race as above, is to assume that there are random productivity differences across firms that are multiplicative in the unobserved productivity of a worker. In that case, the differences in the variances of X_B^{II} and X_W^{II} map directly into unobservables that vary across firms with relative variances proportional to the relative variances of X_B^{II} and X_W^{II}. Alternatively, employers may make expectational errors and rather than assigning a zero expectation to the unobservable assign a random draw based on the distribution of unobservables.

Returning to the main line of argument, the difference between Equations 10 and 10' – the success rates for black and white job applicants – is intended to be informative about discrimination. However, even if $\gamma' = 0$, so there is no discrimination, these two expressions need not be equal because σ_B^{II} and σ_W^{II}, the standard deviations of X_B^{II} and X_W^{II}, can be unequal. The earlier intuition about relative variance of the unobservables and the level of standardization of X^I can be made more precise. Consider the earlier case with $\gamma' = 0$, but $\sigma_W^{II} > \sigma_B^{II}$. Then if X^{I*} is set at a low level – that is, the standardization level is low – characterized by $\beta_I' X^{I*} < c'$, $\sigma_W^{II} > \sigma_B^{II}$ implies that $\Phi[(\beta_I' X^{I*} + \gamma' - c')/\sigma_B^{II}] < \Phi[(\beta_I' X^{I*} - c')/\sigma_W^{II}]$, generating spurious evidence of discrimination against blacks. When $\beta_I' X^{I*} > c'$, we instead get spurious evidence of discrimination in favor of blacks, and switching the relative magnitudes of σ_W^{II} and σ_B^{II} reverses these results.[6]

Thus, even if the means of the unobserved productivity-related variables are the same for each group, and firms use the same hiring standard (that is, $\gamma' = 0$), correspondence studies can

generate evidence consistent with discrimination against blacks (or, alternatively, in their favor). Although demonstrated in the case of normally distributed unobservables, the argument holds for symmetric distributions (Heckman 1998). This is the basis for HS's claim that even under ideal conditions correspondence (or audit) studies are uninformative about discrimination.

IV. Detecting Discrimination

With the right data from a correspondence study, the framework from the preceding section can be used to recover an unbiased estimate of discrimination, conditional on an identifying assumption. The intuition is as follows. The HS critique rests on differences between blacks and whites in the variances of unobserved productivity. The fundamental problem, as Equations 10 and 10' show, is that we cannot separately identify the effect of race (γ') and a difference in the variance of the unobservables ($\sigma_B^{II}/\sigma_W^{II}$). But a higher variance for one group (say, whites) implies a smaller effect of observed characteristics on the probability that a white applicant meets the standard for hiring. Thus, information from a correspondence study on how variation in observable qualifications is related to employment outcomes can be informative about the relative variance of the unobservables, and this, in turn, can identify the effect of discrimination. Based on this idea, the identification problem is solved by invoking an identifying assumption – specifically, that there is variation in some applicant characteristics in the study that affect perceived productivity and have effects that are homogeneous across the races.

More formally, Equations 10 and 10' imply that the difference in outcomes between blacks and whites is

(11) $\Phi[(\beta_I'X^{I*} + \gamma' - c')/\sigma_B^{II}] - \Phi[(\beta_I'X^{I*} - c')/\sigma_W^{II}]$.

In a standard probit, we can only identify the coefficients relative to the standard deviation of the unobservable, so we normalize by setting the variance of the unobservable to

equal one. In this case, impose the normalization for whites only, or $\sigma_W^{II} = 1$. The parameter σ_B^{II} is then the variance of the unobservable for blacks relative to whites. To make this clear, replace σ_B^{II} with $\sigma_{BR}^{II} = \sigma_B^{II}/\sigma_W^{II}$. The normalization $\sigma_W^{II} = 1$ is equivalent to defining all of the coefficients in Equation 11 as their ratios relative to σ_W^{II}. Dropping the prime subscripts to indicate that the coefficients are now defined in relative terms, with this normalization Equation 11 becomes

(11') $\Phi[(\beta_I X^{I*} + \gamma - c)/\sigma_{BR}^{II}] - \Phi[\beta_I X^{I*} - c]$.

As Equation 11' shows, without knowing σ_{BR}^{II} we cannot tell whether the intercepts of the two probits – and hence the hiring probabilities – differ because $\gamma \neq 0$ or because $\sigma_{BR}^{II} \neq 1$. However, if there is variation in the level of qualifications used as controls (X^{I*}), and these qualifications affect hiring outcomes, then we can identify β_I/σ_{BR}^{II} and β_I in Equation 11', and the ratio of these two estimates provides an estimate of σ_{BR}^{II}. This lets us test the hypothesis of equal standard deviations (or variances) of the unobservables. Finally, identification of σ_{BR}^{II} implies identification of γ. Without meaningful variation in X^{I*} (that is, the variation that affects hiring) this is not possible, since in that case all we have in the model are different intercepts with different parameters in both the numerators and the denominators ($(\gamma - c)/\sigma_{BR}^{II}$ and c).

The critical assumption to identify σ_{BR}^{II} and hence γ is that β_I is the same for blacks and whites. Otherwise, the ratio of the two coefficients of X^{I*} for blacks and whites does not identify σ_{BR}^{II}. As HS point out, the constancy of β_I is assumed in the Urban Institute studies that they critique, with discrimination entering through an intercept shift in the evaluation of a worker's productivity, depending on their race. It is not hard to come up with reasons why the coefficients relating X^I to productivity might differ by race. For example, blacks and whites on average attend different schools, and if whites' schools are higher quality, a given number of years of

schooling may do more to increase white productivity than black productivity. But in a correspondence or audit study, it should be possible to control for these kinds of differences; for example, in this case one can control for the area where applicants live, and hence hold school district constant (for example, BM 2004). In other words, in these field experiments the researcher has the capacity to generate data making it more likely that the identifying assumption holds.

HS raise other possibilities. One is that there is discrimination in evaluating particular attributes of a group. For example, employers may discriminate against high-education blacks but not low-education blacks. It is not possible to rule out differences in coefficients arising for these reasons. Finally, HS also suggest that differences in coefficients may reflect "statistical information processing," given incomplete information about productivity, as in statistical discrimination models. Of course this is the idea underlying the identification strategy suggested above, as the difference in β_I for blacks and whites is assumed to reflect precisely the accuracy with which X^I signals productivity for each race. However, as discussed below, when there is data on multiple productivity-related characteristics, there is more one can do to test whether there is homogeneity in the coefficients that allows identification of σ_{BR}^{II} and hence γ.

The estimation of β_I/σ_{BR}^{II} and β_I, and inference on their ratio ($\sigma_{BR}^{II} = \sigma_B^{II}/\sigma_W^{II}$), can be done via a heteroskedastic probit model (for example, Williams 2009), which allows the variance of the unobservable to vary with race. To do this, pool the data for blacks and whites. Similar to Equation 5, letting i denote applicants and j firms, there is a latent variable for perceived productivity relative to the threshold, assumed to be generated by

$$(12) \quad T(P_{ij}^*) = -c + \beta_I X_{ij}^{I*} + \gamma R_i + \varepsilon_{ij}.$$

As is standard, it is assumed that $E(\varepsilon_{ij}) = 0$. But the variance is assumed to follow

(13) $\text{Var}(\varepsilon_{ij}) = [\exp(\mu + \omega R_i)]^2$.

This model can be estimated via maximum likelihood. The observations should be treated as clustered on firms to obtain a variance-covariance matrix that is robust to the dependence of observations across firms. The normalization $\mu = 0$ can be imposed, given that there is an arbitrary normalization of the scale of the variance of one group (in this case whites, with $R_i = 0$). Then the estimate of $\exp(\omega)$ is exactly the estimate of σ_{BR}^{II}.

In this heteroskedastic probit model, the assumption that β_I is the same for blacks and whites identifies γ.[7] Observations on whites identify $-c$ and β_I, and observations on blacks identify $(-c + \gamma)/\exp(\omega)$ and $\beta_I/\exp(\omega)$. Thus, the ratio of $\beta_I/\{\beta_I/\exp(\omega)\}$ identifies $\exp(\omega)$, which, from Equation 13, is the ratio of the standard deviation of the unobservable for blacks relative to whites and is, as before, identified from the ratio of the effect of X^{I*} on blacks relative to its effect on whites. With the estimate of $\exp(\omega)$ (or equivalently σ_{BR}^{II}), along with the estimate of c identified from whites, the expression $(-c + \gamma)/\exp(\omega)$ identified from blacks identifies γ as well.[8]

If $\sigma_{BR}^{II} = 1$, then there is no bias from differences in the distribution of unobservables. Alternatively, if $\sigma_{BR}^{II} \neq 1$, but we had some evidence on how the level of standardization X^{I*} compares to the relevant population of job applicants, we could determine the direction of bias. For example, if the study points to discrimination and there is a bias *against* this finding – based on the estimate of the ratio of variances and information about X^{I*}, then the evidence of discrimination is not spurious, because it would be even stronger absent this bias. But because we can identify γ directly under the assumptions above, we can recover an estimate of discrimination that is not biased by the difference in the variances of the unobservables. And we can do this *without* determining whether X^{I*} used in the study is a high or low level of

standardization, which may be impossible to establish.

The identification of γ and σ_{BR}^{II} depends on the assumption that the unobservables are distributed normally. However, the approach need not be couched solely in terms of normally distributed unobservables and the probit specification. What is required is the separate identification of the effect of race in the latent variable model and the relative variance of the unobservables for blacks and whites. Although typically (for example, Maddala 1983) the logit model is not written with the standard deviation of the error term appearing, it is possible to rewrite it in this way, in which case the difference in coefficients would again be informative about the ratio of the variances of the unobservables (Johnson and Kotz 1970, p. 5).

On the other hand, in the specific setting of this paper – a discrete outcome (hiring) and a discrete treatment (race) – there is no clear way to separately identify how race affects the latent variable and the variance *without* distributional assumptions. This case is covered most explicitly in Manski (1988), who shows that when – in this context – the distribution of the unobservable differs by race, identification of the "structural" coefficient (γ, in this case) requires a parametric specification of the unobservable distribution unless various kinds of continuity assumptions are imposed on the distribution of the observables (with a tradeoff between the tightness of the restrictions on the distribution of the unobservables and on the observables); but the latter do not hold for a discrete treatment.[9] There is a budding literature on identification of variables with dichotomous outcomes with non-parametric or semi-parametric methods. Most prominently, perhaps, Matzkin (1992) shows how this kind of model can be identified non-parametrically in contexts where restrictions from economic theory (concavity, homogeneity of degree one) apply. But these types of restrictions do not apply here.

Moreover, a specific form of the hiring rule is assumed. Different hiring rules are possible, such as one that minimizes the distance between a worker's skill level and the required

skill level on the job (for example, Rothschild and Stiglitz 1982). There is no claim being made that the identification result or the conclusions about discrimination that follow are invariant to different assumptions about the distribution of the unobservables or the hiring rule. Rather than deriving general results, the goal is to show that, in the specific context explored by HS in which they showed that it was impossible to identify discrimination, an additional assumption – which itself has testable implications that are discussed next – permits the identification of discrimination. It remains a question for future research whether it is possible to treat the data from a correspondence study in a less restrictive manner and still learn something about the effects of interest.

The assumption that β_I is the same for blacks and whites cannot be tested if there is only one productivity control. With more controls, however, there is a testable restriction, because if the effects on hiring of multiple productivity controls differ between blacks and whites *only* because of the difference in the variance of the unobservables, the ratios of the estimated probit coefficients for blacks and whites, for each variable, should be the same. Consider the case with two observables X^I and Z^I, modifying Equation 11' to be

(11'') $\Phi[(\beta_I{}^B X^{I*} + \delta_I{}^B Z^{I*} + \gamma - c)/\sigma_{BR}{}^{II}] - \Phi[(\beta_I{}^W X^{I*} + \delta_I{}^W Z^{I*} - c)]$,

where the coefficients on the observables have B and W superscripts to denote possible differences by race. Normalize the coefficients in the first expression so that $\tilde{\beta}_I{}^B = \beta_I{}^B/\sigma_{BR}{}^{II}$ and $\tilde{\delta}_I{}^B = \delta_I{}^B/\sigma_{BR}{}^{II}$. If the coefficients in Equation 11'' do not differ by race, but only the variances of the unobservables differ, then

(14) $\tilde{\beta}_I{}^B/\beta_I{}^W = \tilde{\delta}_I{}^B/\delta_I{}^W$.

In other words, the black/white ratios of the coefficients from separate probits for blacks and whites, or from a probit with a full set of race interactions, differ from 1 only because $\sigma_{BR}{}^{II} \neq$

1, and hence should be equal. Thus, the restrictions implied by homogeneity of effects but unequal variances of the unobservables can be tested. Of course failure to reject the restrictions does not decisively rule out the possibility that $\sigma_{BR}^{II} = 1$, with the coefficients differing by race for other reasons but Equation 14 still holding. With a larger number of control variables, however, it seems unlikely that this alternative scenario would explain failure to reject the restrictions in Equation 14. It is also possible to choose – as an identifying assumption – a subset of the observable characteristics for which Equation 14 holds, and to identify σ_{BR}^{II} only from the coefficients of this subset of variables, or to do this and then test the restriction on the other coefficients as overidentifying restrictions.

A final issue concerns the interpretation of the coefficients from the heteroskedastic probit model. Consider a model with generic notation, where the latent variable depends on a vector of variables S and coefficients ψ, and the variance depends on a vector of variables T, which includes S, with coefficients θ. The elements of S are indexed by k. For a standard probit, coefficient estimates are translated into estimates of the marginal effects of a variable using

(15) $\partial P(\text{hire})/\partial S_k = \psi_k \phi(S\psi)$,

where S_k is the variable of interest with coefficient ψ_k, $\phi(.)$ is the standard normal density, and the standard deviation of the unobservable is normalized to one. Typically this is evaluated at the means of S. When S_k is a dummy variable such as race, the difference in the cumulative normal distribution functions is often used instead, although the difference is usually trivial.

The marginal effect is more complicated in the case of the heteroskedastic probit model, because if the variances of the unobservable differ by race, then when race "changes" both the variance and the level of the latent variable that determines hiring can shift. As long as we use the continuous version of the partial derivative to compute marginal effects from the heteroskedastic probit model, there is a natural decomposition of the effect of a change in a

variable S_k that also appears in T into these two components. In particular, generalize the notation of Equation 13 to

(13') $\text{Var}(\varepsilon) = [\exp(T\theta)]^2$,[10]

with the variables in T arranged such that the k^{th} element of T is S_k. Then the overall partial derivative of P(hire) with respect to S_k is

(16) $\partial P(\text{hire})/\partial S_k = \phi(S\psi/\exp(T\theta))\cdot\{(\psi_k - S\psi\cdot\theta_k)/\exp(T\theta)\}$.[11]

This expression can be broken into two pieces. First, the partial derivative with respect to changes in S_k affecting only the level of the latent variable – corresponding to the counterfactual of S_k changing the valuation of the worker without changing the variance of the unobservable – is equal to

(16') $\phi(S\psi/\exp(T\theta)) \cdot\{\psi_k/\exp(T\theta)\}$.

Second, the partial derivative with respect to changes via the variance of the unobservable is equal to

(16'') $\phi(S\psi/\exp(T\theta))\cdot\{(-S\psi\cdot\theta_k)/\exp(T\theta)\}$.

In the analysis below, these two separate effects are reported as well as the overall marginal effect, and standard errors are calculated using the delta method. The effect of race via how it shifts the latent variable – or how race shifts the employer's valuation of worker productivity – is of greatest interest. The point of the HS critique is that differential treatment of blacks and whites based only on differences in variances of the unobservable should not be interpreted as discrimination. And moreover, as argued in Section VI, the effect of race via the latent variable captures discrimination likely to be manifested in the real economy, whereas its effect through the variance is more of an artifact of the study.

V. Evidence, Implementation, and Assessment

A. Existing Evidence

As the preceding discussion shows, we need information on the effects of productivity-related characteristics on hiring or callbacks, estimated separately for blacks and whites (or other groups), to identify discrimination in an audit or correspondence study, or even to assess likely biases. Reporting of such evidence is rare in the literature, because these studies typically create one "type" of applicant for which there is only random variation in characteristics that does not (or is not intended to) affect outcomes. However, BM's well-known correspondence study of race discrimination is unusual in that – for reasons unrelated to the concerns of this paper – it uses two types of applicants.[12]

Part of their analysis studies callback differences by race for resumes that they constructed to be low versus high quality, to ask whether blacks and whites have different incentives to invest in skills, as in Lundberg and Startz (1983). White callback rates are higher for both types of resumes. But although white callback rates increase significantly with resume quality (from 8.5 to 10.8 percent), black callback rates increase only slightly (from 6.2 to 6.7 percent) and the change is not statistically significant. Similar qualitative conclusions are reached based on an analysis that measures resume quality for one part of the sample based on the predicted probability of callbacks estimated from another part of the sample. In this analysis, both groups experience an increase in callback rates from higher-quality resumes, but the effect is larger for whites.

Similarly, BM report probit models estimated for whites and blacks separately (their Table 5). These estimates reveal substantially stronger effects of measured qualifications for whites than for blacks. Among the estimated coefficients that are statistically significant for at least one group, effects are larger for whites for experience,[13] having an email address, working

while in school, academic honors, and other special skills (such as language). The only exception is for computer skills, which inexplicably have a negative effect on callback rates for whites.[14]

As the present paper suggests, an alternative interpretation of smaller estimated probit coefficients or marginal effects for blacks than for whites is a difference in the variance of the unobservables. In particular, the lower coefficients for blacks are consistent with a larger variance for blacks, or $\sigma_{BR}^{II} > 1$. If it is also true that BM standardized applicants at low levels of the control variables, then the HS analysis would imply that there is a bias towards finding discrimination *in favor of* blacks; that is, the evidence of discrimination against blacks would be even stronger absent the bias from differences in the distribution of unobservables. BM explicitly state that they tried to avoid overqualification even of the *higher-quality* resumes (p. 995). But it is very difficult to assess whether the characteristics of applicants were low, since there is no way to identify the population of applicants. Hence, implementation of the estimation procedure proposed in this paper is likely the only way even to *sign* the bias, let alone to recover an unbiased estimate of discrimination.

B. Implementation Using Bertrand and Mullainathan Data

Because BM's data include applicants with different levels of qualifications, and the qualifications predict callbacks, their data can be used to implement the methods described above. Table 1 begins by simply presenting probit estimates for the probability of a callback. Marginal effects are reported for specifications with no controls except a dummy variable for females, adding controls for the individual characteristics included on the resumes, and finally adding also neighborhood characteristics for the applicant's zip code; the specific variables are listed in the footnote to the table. Estimates are shown for males and females combined, and for females only; as the sample sizes indicate, the male sample is considerably smaller.[15] Aside

from the estimated effects of race, estimates are shown for a few of the resume characteristics capturing applicants' qualifications.

Echoing BM's conclusions, there is a sizable and statistically significant difference between the callback rates for blacks and whites, with the rate for blacks lower by 3-3.3 percentage points (or about 33 percent relative to the white callback rate of 9.65 percent). The estimated race differences are robust to the inclusion of the different sets of control variables, which is what we should expect since the resume characteristics are assigned randomly.

Interestingly, in light of the results of other audit and correspondence studies, there is no evidence of lower callback rates for females than for males. And although not reported in the tables, this was true if the same methods used below to recover unbiased estimates of race discrimination were applied to the estimation of sex discrimination. However, BM's study was to a large extent focused on jobs typically held by females, and was not designed to test for sex discrimination.[16] Table 1 also shows that a number of the resume characteristics have statistically significant effects on the callback probability; this, of course, is an essential input for using the methods described above to recover an unbiased estimate of discrimination.

The main analysis is reported beginning in Table 2, for the specifications with the full set of individual resume controls, and then adding as well the full set of neighborhood controls. Panel A simply repeats the estimated race effects from Table 1, for comparison. Panel B begins by reporting the estimated overall marginal effects of race from the heteroskedastic probit model (Equation 16). As the table shows, these estimates are slightly smaller (in absolute value) than the estimates from the simple probits. They remain statistically significant and indicate callback rates that are lower for blacks by about 2.4-2.5 percentage points (or about 25 percent).

Decomposing the marginal effect, the effect via the level of the latent variable is larger than the marginal effect from the probit estimation, ranging from −0.054 to −0.086. The effect

of race via the variance of the unobservable, in contrast, is positive, ranging from 0.028 to 0.062. (This latter effect is not statistically significant.) The implication is that race discrimination is more severe than indicated by the analysis that ignores differences in the variances of the unobservables. The evidence that the probit estimates understate discrimination against blacks is consistent with a low level of standardization of X^{I*}, coupled with a higher estimated variance of the unobservable for blacks, as conjectured earlier based on BM's results. And as reported in the next row of the table, the estimated ratio of the standard deviation of the unobservable for blacks to the standard deviation for whites always exceeds one, although the difference is not statistically significant. The positive effect of being black via the variance is what we expect if X^{I*} is low, since then a larger relative variance for blacks increases the relative probability that they are hired (called back).

The next two rows of the table report diagnostic test statistics. First, the p-values from the test of the overidentifying restrictions (Equation 14) are shown, based on probit specifications interacting all of the controls with race. In all four cases the restrictions are not rejected, with p-values ranging from 0.17 to 0.62. Nonetheless, the lower end of this range of p-values suggests that the restrictions sometimes might be fairly inconsistent with the data. As a consequence, below some alternative estimates are discussed that use only a subset of variables, for which Equation 14 is more consistent with the data, to identify discrimination.

Finally, the subset of control variables for which the absolute value of the estimated coefficient for whites exceeded that for blacks – consistent with the larger standard deviation of unobservables for blacks – was identified. Then the heteroskedastic probit model was estimated leaving the race interactions of the other variables in the model – so that the restrictions from Equation 14 that were less consistent with the data were not imposed – and the joint significance of these latter variables was tested. Despite this latter subset of variables having estimated

coefficients less consistent with the restrictions in Equation 14, the p-values indicate that these interactions can also be excluded from the model. This can be viewed as an overidentifying test of the restriction that there are no differences in the effects of any of the control variables by race, for the specifications for which the estimates are reported in the first row of Panel B. Although technically it is only necessary to assume that there is a single variable for which the coefficient is the same for blacks and whites, there is no obvious variable to choose for the purposes of identification; here, instead, I let the data select a set of variables more consistent with the identifying restriction.

Table 3 follows up on the last procedure, by instead simply dropping from the analysis the control variables for which the absolute value of the estimated coefficient for whites was less than for blacks. Given that the control variables are random with respect to race, dropping controls does not introduce bias. As we would expect, the p-values for the tests of this set of restrictions are now much closer to one, ranging from 0.68 to 0.92, compared with a range of 0.17 to 0.62 in Table 2. However, as the table shows, the estimated effects of race are similar to those in Table 2 and do not point to any different conclusions.

C. Monte Carlo Assessment

A Monte Carlo assessment of how well the estimation procedure proposed in this paper works in terms of removing the bias in estimates of discrimination from correspondence study evidence was carried out. The Monte Carlo assessment uses simulated data of the type needed to implement the estimator – namely, data with applicants at two different levels of productivity. The analysis is described in detail in Appendix 1; here the findings are briefly summarized.

First, using simulated data either with or without discrimination, the heteroskedastic estimation procedure eliminates the bias, and recovers an unbiased estimate of discrimination. (And as a benchmark, the simulated data are used to replicate the HS critique, showing that a

simple probit analysis can generate bias in any direction.) Second, for a specific case we consider, when the identifying assumption is violated and the coefficient on the productivity-related characteristic in the simulated data is not equal for blacks and whites, but is treated as equal in the estimation, there are two findings. When there is no discrimination, the misspecification has no effect; the estimator still produces estimates of γ centered on zero. When there is discrimination and the model is misspecified, there is bias but it is multiplicative, so the estimator will not generate the wrong sign for the estimate of γ.

VI. The Meaning of Discrimination

The fact that a correspondence study can generate evidence of discrimination when $\gamma = 0$ raises the question of whether the evidence reflects a different kind of discrimination. In this case, the productivity of blacks and whites are regarded equally by employers (or equivalently there is no taste discrimination). Moreover, employers are not making any assumption about mean differences in unobservables between blacks and whites. However, they are making assumptions about distributional differences with regard to the variance of unobservables, and it is these assumptions that lead them, given the level of standardization of the study applicants, to prefer blacks or whites – which might be labeled "second-moment" statistical discrimination.

The HS critique can be recast as showing that the analysis of data from a standard audit or correspondence study cannot distinguish between discrimination as it usually interpreted, and discrimination based on different variances of unobservables. Indeed Manski's (1988) paper discussed earlier, in the context of identification, makes this same point, noting that we can estimate the "reduced-form" effect of a binary treatment variable on a binary outcome without strong assumptions, but not the "structural" effect. The present paper imposes an assumption to identify the structural parameter γ, distinguishing between what is typically viewed as discrimination (stemming from tastes or, as noted earlier, from standard statistical

discrimination) and different treatment stemming from differences in variances of the unobservable.

A natural question, then, is whether the structural effect of race, captured in γ, is of interest, or whether instead all we want to know is the reduced-form effect of race – that is, how race affects the probability of hiring *whether* because employers discount black workers' productivity (for example) or because employers treat blacks and whites differentially because of different distributions of the unobservable. There are two reasons why the structural coefficient γ is important. First, to the best of my knowledge, differential treatment based on assumptions (true or not) about variances are not viewed as discriminatory in the legal literature.[17] Thus, identification of γ speaks to the discrimination that is most clearly illegal.

The second and more compelling argument is that taste discrimination or "first-moment" statistical discrimination, captured in γ, generalizes from the correspondence study to the real economy. In contrast, the second-moment statistical discrimination is an artifact of how a correspondence study is done – in particular, the standardization of applicants to particular, and similar, values of the observables, relative to the actual distribution of observables among real applicants to these firms.

Suppose that the actual population of applicants to the employers included in a correspondence study comes from a large range of values of X^I. Then the different distributions of the unobservables by race can have strong effects on hiring outcomes in the correspondence study because in the study the applicants are standardized to a narrow range of X^I. In that sense, the differential treatment by race is strongly influenced by the design of the correspondence study, rather than by behavior of real firms evaluating real applicants, and can be generated *solely* from the standardization of applicants in a narrow range of X^I; in that sense, "discrimination" attributable to differences in the variance of the unobservables is an artifact of

the study.[18] In contrast, the structural effect of race via the latent variable would generalize to how the firms in the correspondence study, and presumably similar firms, evaluate actual job applicants and make hiring decisions.

VII. Conclusions and Discussion

Many researchers view audit and correspondence studies as the most compelling way to test for labor market discrimination. And research applying these methods to many different types of groups nearly always finds evidence of discrimination. The use of audit studies to test for labor market discrimination has been criticized on numerous grounds having to do with whether applicants from different groups appear identical to employers. Many of these criticisms can be countered by using correspondence studies of fictitious applicants on paper rather than fictitious in-person applicants.

However, Heckman and Siegelman (1993) show that even in correspondence studies in which group averages are identical conditional on the controls, group differences in the variances of unobservable dimensions of productivity can invalidate the empirical tests, leading to spurious evidence of discrimination in either direction, or spurious evidence of an absence of discrimination. This is a fundamental criticism of correspondence studies, as it implies that evidence regarding discrimination from even the best-designed correspondence study can be misleading. Nonetheless, this criticism has been ignored in the literature.

This paper shows that if a correspondence study includes observable measures of variation in applicants' quality that affect hiring outcomes, an unbiased estimate of discrimination can be recovered even when there are group differences in the variances of the unobservable. The method is applied to Bertrand and Mullainathan's (2004) correspondence study, and leads to stronger evidence of race discrimination that adversely affects blacks than is obtained when differences in the variances of the unobservable are ignored. Moreover, this

conclusion is bolstered by Monte Carlo simulations suggesting that the estimation procedure performs well, eliminating the problems highlighted by Heckman and Siegelman that could otherwise lead to badly misleading conclusions from the analysis of data from correspondence (or audit) studies.

Finally, it should be recognized that the method proposed here can be easily implemented in any future correspondence (or audit) study. All that is needed is for the resumes or applicants to include some variation in characteristics that affect the probability of being hired.[19] This is different from what is often done in designing these studies, where researchers try to create a pool of equally-qualified applicants. All that needs to be done is to intentionally create resumes of different quality. Once a researcher confirms that a set of productivity-related characteristics on the resumes affected hiring outcomes, it should then be possible – conditional on an identifying assumption that has testable implications – to detect discrimination.

Appendix 1. Monte Carlo Assessment

This appendix provides Monte Carlo evidence on how well the estimation procedure proposed in this paper works in terms of removing the bias in estimates of discrimination from correspondence study evidence, and explores the consequences of violation of the identifying assumption. Figure A1 replicates the basic result from Heckman (1998), showing that probit analysis of the data from a correspondence study can generate substantial bias in either direction. Paralleling Heckman, this is done for the case in which $c = 0$, $\beta_I = 1$, $Var(X_W^{II})/Var(X_B^{II}) = (\sigma_W^{II})^2/(\sigma_B^{II})^2 = 2.25$,[20] and there is no discrimination ($\gamma = 0$). For the Monte Carlo simulations, the assumed data generating process is $X^{I*} \sim N(0,1)$, $X_B^{II} \sim N(0,1)$, $X_W^{II} \sim N(0,2.25)$. Paralleling the standardization of correspondence study applicants, the data are generated by sampling X^{I*} from a truncated normal distribution, in steps of $0.1 \pm 0.1 \cdot SD(X^{I*})$. The simulation is done 100 times at each value of X^{I*} shown in the graph, with samples of 2,000 blacks and 2,000 whites in

each simulation (roughly BM's sample sizes), and a probit model is estimated for each simulated data set. The panels in the figure show – for both the estimates of γ and the marginal effects – the true values based on the assumed parameters, and the means based on the estimates.[21]

The upper figures clearly illustrate that, despite the absence of discrimination in the data generating process (the true effect is constant at 0), the evidence can either point to discrimination against blacks or discrimination in favor of blacks, depending on the level of standardization of X^{I*}. The marginal effects show that even though there is no discrimination in the data generating process, quite strong evidence of discrimination in either direction can emerge, with a marginal effect of −0.1 (0.1) for low (high) values of X^{I*}. Finally, as we would expect, only at $X^{I*} = 0$ is the estimate of γ (and the marginal effect) unbiased. The lower panels of Figure A1 report the same kind of evidence when $\gamma = -0.5$, consistent with discrimination. A similar result is apparent, with substantial bias relative to the true γ or the true marginal effect.

The heteroskedastic probit estimation requires data with multiple levels of the value of X^{I*} at which applicants are standardized. As an intermediate step to separate the consequences of generating the data this way, and the consequences of implementing the heteroskedastic probit estimator, Figure A2 shows results with such generated data, but continuing to use the probit specification. X^{I*} is now sampled from two truncated normal distributions, one using X^{I*}in steps of $0.1 \pm 0.1 \cdot SD(X^{I*})$, as before, and the second using instead $X^{I*} + 0.5$, again in steps of $0.1 \pm 0.1 \cdot SD(X^{I*})$. Figure A2 shows qualitatively similar results to Figure A1, so simply using data with variation in productivity-related characteristics does not, in itself, eliminate the bias. Nonetheless, the biases in both the no discrimination and discrimination cases are a bit smaller than in Figure A1 because of the larger range covered by X^{I*}.[22]

Figure A3 reports results for the heteroskedastic probit estimation, using the same data generating process for simulating data as in Figure A2, although in this case 5,000 simulations

are run for each pair of values of X^{I*} because the heteroskedastic probit estimation is less precise than the simple probit estimation. The top panel covers the no discrimination case ($\gamma = 0$). The left-hand graph shows the means of the true and estimated values of the marginal effects for each value of X^{I*}. These are largely indistinguishable in the figure, indicating no bias. The right-hand panel provides evidence on the distribution of the estimates, showing the distance between the 25th and 75th percentiles of the estimates and between the 2.5th and 97.5th percentiles at each value of X^{I*}. The distribution of estimates is quite tight at levels of standardization near the center of the distribution of X^{I*}, but becomes wider at more extreme values, when hiring rates in the generated data move towards zero or one. The discrimination case ($\gamma = -0.5$) similarly demonstrates that the heteroskedastic probit estimation eliminates the bias.

The last analysis, reported in Figure A4, considers the implications of the data generating process violating the identifying assumption that the coefficient(s) on the productivity-related characteristics are equal for blacks and whites. Results are presented for two cases: mild violation in which the coefficient on X^{I*} (β_I) is slightly larger for whites than for blacks (1.1 versus 1); and strong violation in which it is much larger (2 versus 1). As Figure A4 shows, in the case of no discrimination – the left-hand panels – the results are indistinguishable from when the identifying assumption is not violated. In contrast, in the discrimination case the estimated marginal effects become more negative than the true effects over much of the range, only slightly with mild violation of the identifying assumption, but more so when the violation is more pronounced.

The implications of what happens when the identifying assumption is violated in this specific setting make sense, thinking about how γ is identified. Using estimates of the separate probits in Equations 10 and 10', the ratio of the standardized white probit coefficient to the black probit coefficient identifies σ_{BR}^{II} (which equals $\sigma_B^{II}/\sigma_W^{II}$). When the true value of β_I is larger for

whites than for blacks, but it is assumed that they are equal, σ_{BR}^{II} is overestimated. For example, in the case in the top panel of Figure A4, the ratio of coefficients is $(\beta_I \cdot 1.1)/(\beta_I/\sigma_{BR}^{II}) = 1.1 \cdot \sigma_{BR}^{II}$. Recall from the earlier discussion that the probit for blacks identifies $(-c + \gamma)/\exp(\omega) = (-c + \gamma)/\sigma_{BR}^{II}$. Because $c = 0$ in the simulations, we identify γ by multiplying the estimate of this expression by the estimate of σ_{BR}^{II}; the upward bias in the estimate of σ_{BR}^{II} therefore implies that the estimate of γ is biased away from zero. In the no discrimination case, when $\gamma = 0$, this is irrelevant; multiplying an estimate that averages zero by the upward-biased estimate of σ_{BR}^{II} has no effect. But when the true γ is non-zero (and negative), this bias leads to an estimate of γ that is more negative. When γ is more negative, we get exactly the "bending" of the estimated marginal effects that the right-hand panels of Figure A4 illustrate.[23] A violation of the assumption in the opposite direction (β_I larger for blacks) would lead to biases in the opposite direction. Nonetheless, it follows from this reasoning that the bias is multiplicative, and hence does not generate the wrong sign for the estimate of γ, or generate spurious evidence of discrimination when there is no discrimination. However, further analysis shows that when $c \neq 0$ (or, more generally, when c is not equal to the expected value of unobserved productivity), the implications of violation of the identifying assumption are less sharp.

References

Aigner, Dennis J., and Glen Cain. 1977. "Statistical Theories of Discrimination in Labor Markets." *Industrial and Labor Relations Review* 30(2):175-87.

Banerjee, Abhijit, Marianne Bertrand, Saugato Datta, and Sendhil Mullainathan. 2008. "Labor Market Discrimination in Delhi: Evidence from a Field Experiment." *Journal of Comparative Economics* 38(1):14-27.

Becker, Gary S. 1971. *The Economics of Discrimination, Second Edition*. Chicago: University of Chicago Press.

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.

Bovenkerk, F., M. Gras, and D. Ramsoedh. 1995. "Discrimination Against Migrant Workers and Ethnic Minorities in Access to Employment in the Netherlands." International Migration Papers, No. 4. Geneva, Switzerland: International Labour Office.

Cornelißen, Thomas. 2005. "Standard Errors of Marginal Effects in the Heteroskedastic Probit Model." Institute of Quantitative Economic Research, Discussion Paper No. 230. Hanover, Germany: University of Hanover.

Dickinson, David L., and Ronald L. Oaxaca. 2009. "Statistical Discrimination in Labor Markets: An Experimental Analysis." *Southern Economic Journal* 71(1):16-31.

Fix, Michael, and Raymond Struyk. 1993. *Clear and Convincing Evidence: Measurement of Discrimination in America*. Washington, DC: The Urban Institute Press.

Gneezy, Uri, and John A. List. 2004. "Are the Disabled Discriminated Against in Product Markets? Evidence from Field Experiments." Unpublished paper. Chicago: University of

Chicago.

Goldberg, Pinelopi Koujianou. 1996. "Dealer Price Discrimination in New Car Purchases: Evidence from the Consumer Expenditure Survey." *Journal of Political Economy* 104(3):622-54.

Heckman, James J. 1998. "Detecting Discrimination." *Journal of Economic Perspectives* 12(2):101-16.

Heckman, James, and Peter Siegelman. 1993. "The Urban Institute Audit Studies: Their Methods and Findings." In Fix and Struyk, eds., *Clear and Convincing Evidence: Measurement of Discrimination in America*. Washington, D.C.: The Urban Institute Press, pp. 187-258.

Johnson, Norman L., and Samuel Kotz. 1970. *Continuous Univariate Distributions – 2*. New York: John Wiley and Sons.

Lahey, Joanna N. 2008. "Age, Women, and Hiring: An Experimental Study." *Journal of Human Resources* 43(1):30-56.

Lahey, Joanna N., and Ryan A. Beasley. 2009. "Computerizing Audit Studies." *Journal of Economic Behavior & Organization* 70(3):508-14.

Lundberg, Shelly J., and Richard Startz. 1983. "Private Discrimination and Social Intervention in Competitive Labor Markets." *American Economic Review* 73(3):340-7.

Maddala, G.S. 1983. *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge, U.K.: Cambridge University Press.

Manski, Charles F. 1988. "Identification of Binary Response Models." *Journal of the American Statistical Association* 83(403):729-38.

Matzkin, Rosa L. 1992. "Nonparametric and Distribution-Free Estimation of the Binary Threshold Crossing and the Binary Choice Models." *Econometrica* 60(2):239-70.

Mincy, Ronald. 1993. “The Urban Institute Audit Studies: Their Research and Policy Context.” In Fix and Struyk, eds., *Clear and Convincing Evidence: Measurement of Discrimination in America*. Washington, DC: The Urban Institute Press, pp. 165-86.

Neumark, David. 1996. “Sex Discrimination in Restaurant Hiring: An Audit Study.” *Quarterly Journal of Economics* 111(3):915-41.

Neumark, David. 1999. “Wage Differentials by Race and Sex: The Roles of Taste Discrimination and Labor Market Information.” *Industrial Relations* 38(3):414-45.

Pager, Devah. 2007. “The Use of Field Experiments for Studies of Employment Discrimination: Contributions, Critiques, and Directions for the Future.” *The Annals of the American Academy of Political and Social Science* 609(1):104-33.

Riach, Peter A., and Judith Rich. 2002. “Field Experiments of Discrimination in the Market Place.” *The Economic Journal* 112(483):F480-518.

Riach, Peter A., and Judith Rich. 2007. “An Experimental Investigation of Age Discrimination in the Spanish Labor Market.” IZA Discussion Paper No. 2654. Bonn, Germany: Institute for the Study of Labor (IZA).

Rothschild, Michael, and Joseph E. Stiglitz. 1982. “A Model of Employment Outcomes Illustrating the Effect of the Structure of Information on the Level and Distribution of Income.” *Economics Letters* 10(3-4):231-6.

Smeeters, B., and A. Nayer. 1998. “La Discrimination a l’Acces a l’Emploi en Raison de l’Origine Etrangere: le Cas de le Belgique.” International Migration Papers, No. 23. Geneva, Switzerland: International Labour Office.

Turner, Margery, Michael Fix, and Raymond Struyk. 1991. “Opportunities Denied, Opportunities Diminished: Racial Discrimination in Hiring.” UI Report 91-9. Washington,

DC: The Urban Institute.

U.S. Equal Employment Opportunity Commission. 1996. Notice 915.002, May 22, http://www.eeoc.gov/policy/docs/testers.html (viewed February 28. 2010).

U.S. Equal Employment Opportunity Commission, n.d. "Facts About Race/Color Discrimination," http://www.eeoc.gov/facts/fs-race.pdf (viewed March 23, 2009).

Williams, Richard. 2009. "Using Heterogeneous Choice Models to Compare Logit and Probit Coefficients Across Groups." Unpublished manuscript. South Bend, Indiana: Notre Dame University.

Table 1
Probit Estimates for Callbacks: Basic Results

	Males and females			Females		
	(1)	(2)	(3)	(4)	(5)	(6)
Black	-.033	-.030	-.030	-.033	-.030	-.030
	(.006)	(.006)	(.006)	(.008)	(.007)	(.007)
Female	.009	-.001	.001	…	…	…
	(.012)	(.011)	(.011)			
Selected individual resume controls						
Bachelor's degree		.009	.009		.019	.019
		(.009)	(.009)		(.010)	(.010)
Experience $\cdot 10^{-1}$		.080	.076		.080	.076
		(.029)	(.028)		(.034)	(.033)
Experience^2 $\cdot 10^{-2}$		-.022	-.021		-.019	-.018
		(.011)	(.010)		(.013)	(.012)
Academic honors		.039	.040		.026	028
		(.015)	(.015)		(.017)	(.017)
Special skills		.056	.055		.060	.059
		(.009)	(.009)		(.010)	(.010)
Other controls:						
Individual resume characteristics		X	X		X	X
Neighborhood characteristics			X			X
Mean callback rate	.080	.080	.080	.082	.082	.082
N	4,784	4,784	4,784	3,670	3,670	3,670

Note: Marginal effects using Equation 15 are reported. Standard errors are computed clustering on the ad to which the applicants responded, and are reported in parentheses; the delta method is used to compute standard errors for the marginal effects. Individual resume characteristics include bachelor's degree, experience and its square, volunteer activities, military service, having an email address, gaps in employment history, work during school, academic honors, computer skills, and other special skills. Neighborhood characteristics include the fraction high school dropout, college graduate, black, and white, as well as log median household income, in the applicant's zip code.

Table 2
Heteroskedastic Probit Estimates for Callbacks: Full Specifications

	Males and females		Females	
	(1)	(2)	(3)	(4)
A. Estimates from basic probit (Table 1)				
Black	-.030	-.030	-.030	-.030
	(.006)	(.006)	(.007)	(.007)
B. Heteroskedastic probit model				
Black (unbiased estimates)	-.024	-.026	-.026	-.027
	(.007)	(.007)	(.008)	(.008)
Marginal effect of race through level	-.086	-.070	-.072	-.054
	(.038)	(.040)	(.040)	(.040)
Marginal effect of race through variance	.062	.045	.046	.028
	(.042)	(.043)	(.045)	(.044)
Standard deviation of unobservables, black/white	1.37	1.26	1.26	1.15
Wald test statistic, null hypothesis that ratio of standard deviations = 1 (p-value)	.22	.37	.37	.56
Wald test statistic, null hypothesis that ratios of coefficients for whites relative to blacks are equal, fully interactive probit model (p-value)	.62	.42	.17	.35
Test overidentifying restrictions: include in heteroskedastic probit model interactions for variables with |white coefficient| < |black coefficient|, Wald test for joint significance of interactions (p-value)	.83	.33	.34	.56
Number of overidentifying restrictions	3	6	2	6
Other controls:				
Individual resume characteristics	X	X	X	X
Neighborhood characteristics		X		X
N	4,784	4,784	3,670	3,670

Note: See notes to Table 1. In the first row of Panel B the marginal effects in Equation 16 are reported, with the decomposition in Equations 16' and 16'' immediately below; the marginal effects are evaluated at sample means. The standard errors for the two components of the marginal effects are computed using the delta method. Test statistics are based on the variance-covariance matrix clustering on the ad to which the applicants responded. Individual resume characteristics also include the variables listed separately in Table 1.

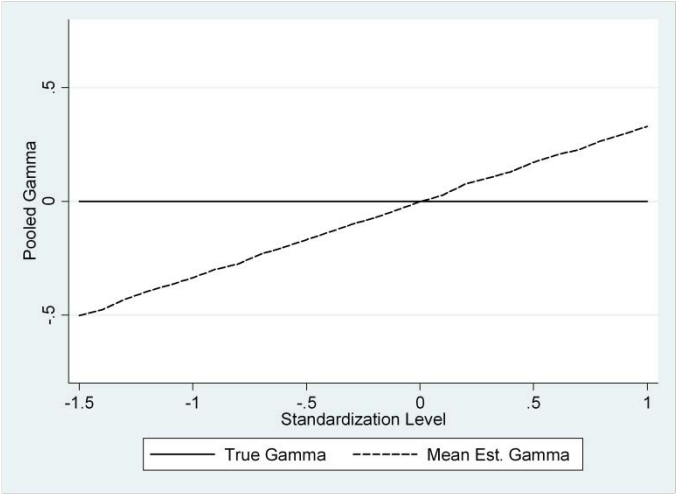
Table 3
Heteroskedastic Probit Estimates for Callbacks: Restricted Specifications Using only Controls with Absolute Value of Estimated Effect in Fully Interactive Probit Model Larger for Whites than Blacks

	Males and females		Females	
	(1)	(2)	(3)	(4)
A. Estimates from basic probit				
Black	-.030	-.030	-.030	-.030
	(.006)	(.006)	(.007)	(.006)
B. Heteroskedastic probit model				
Black (unbiased estimates)	-.024	-.025	-.024	-.025
	(.007)	(.007)	(.009)	(.008)
Marginal effect of race through level	-.090	-.080	-.086	-.077
	(.037)	(.036)	(.040)	(.038)
Marginal effect of race through variance	.066	.056	.062	.052
	(.041)	(.039)	(.044)	(.042)
Standard deviation of unobservables, black/white	1.41	1.33	1.37	1.30
Wald test statistic, null hypothesis that ratio of standard deviations = 1 (p-value)	.19	.23	.25	.29
Wald test statistic, null hypothesis that ratios of coefficients for whites relative to blacks are equal, fully interactive probit model (p-value)	.84	.92	.68	.74
Other controls:				
Individual resume characteristics	X	X	X	X
Neighborhood characteristics		X		X
N	4,784	4,784	3,670	3,670

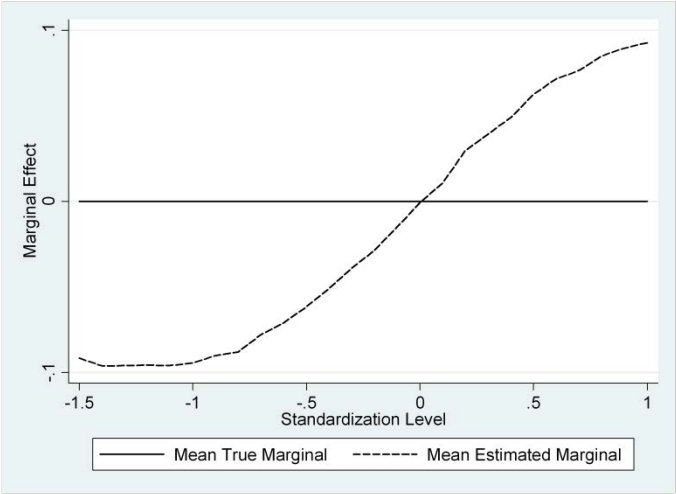
Notes: See notes to Tables 1 and 2.

No discrimination (γ = 0)

Probit estimates of γ

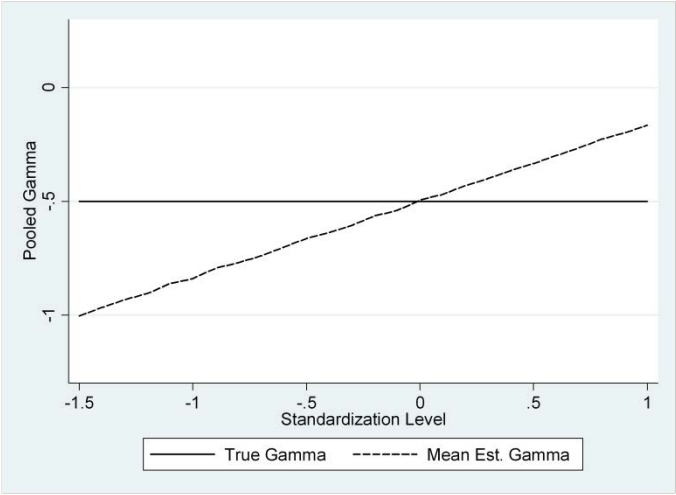


Probit estimates of marginal effect of black

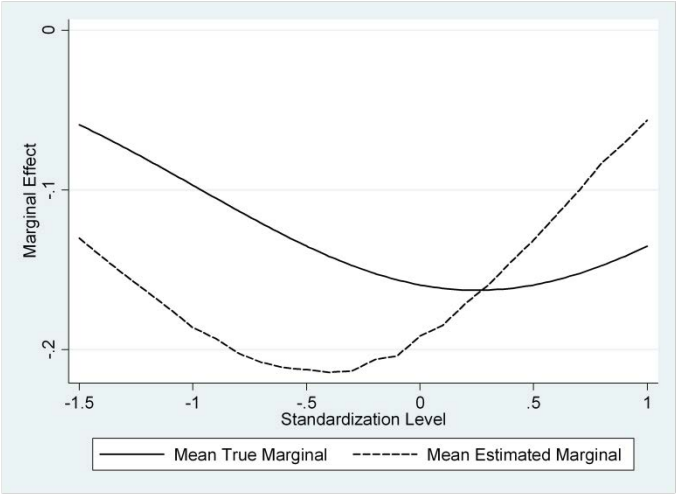


Discrimination (γ = -.5)

Probit estimates of γ



Probit estimates of marginal effect of black



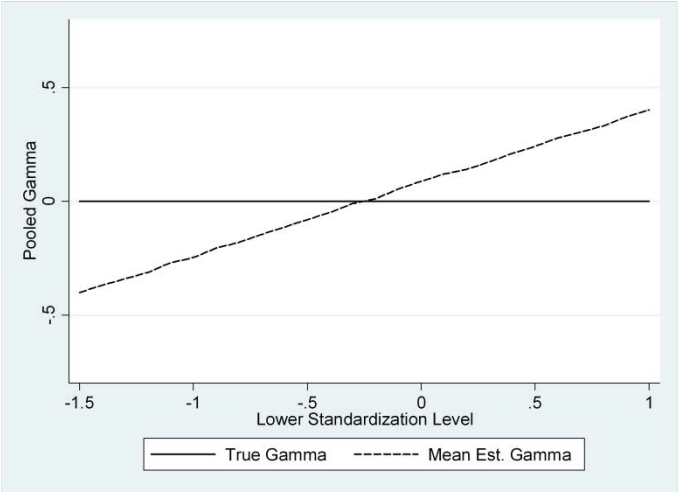
Notes: Left-hand graphs shows true γ and mean estimated γ. Right-hand graphs show marginal effects, evaluated at sample means for simulated data. In the data generating process, $X^{I*} \sim N(0,1)$, $X_B^{II} \sim N(0,1)$, $X_W^{II} \sim N(0,2.25)$, so $Var(X_W^{II})/Var(X_B^{II}) = 2.25$ (X_W^{II} and X_B^{II} are unobservable); $\beta_I = 1$ and $c = 0$ for both blacks and whites. Estimates are generated by Monte Carlo simulation, drawing 4,000 observations (2,000 white and 2,000 black) from truncated normal distribution at each value of X^{I*} (in steps of $0.1 \pm 0.1 \cdot SD(X^{I*})$) and estimating probit model. Simulation is done 100 times at each value of X^{I*}.

Figure A1

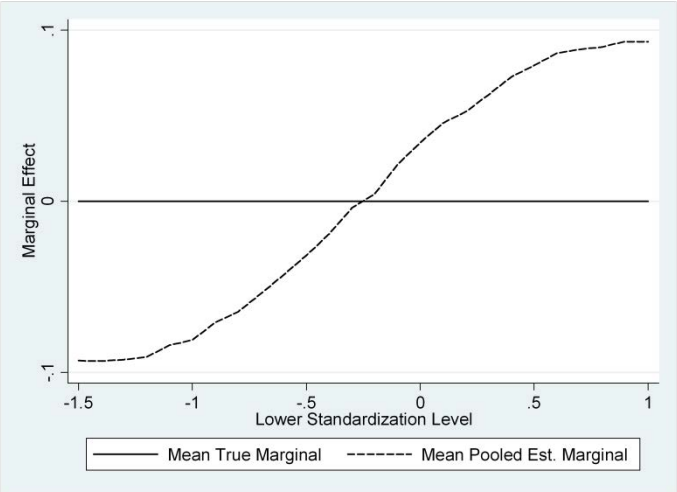
Replication of Heckman (Figure 1, 1998), and Monte Carlo Simulations of Estimates of Marginal Effects from Simple Probit Estimation

No discrimination (γ = 0)

Probit estimates of γ

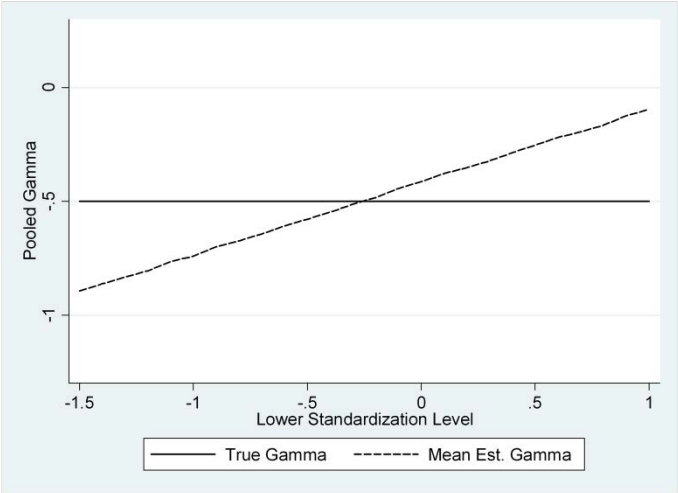


Probit estimates of marginal effect of black

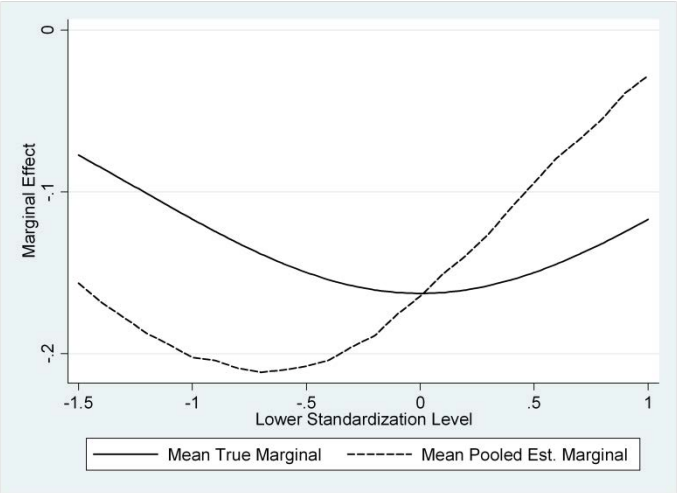


Discrimination (γ = -.5)

Probit estimates of γ



Probit estimates of marginal effect of black

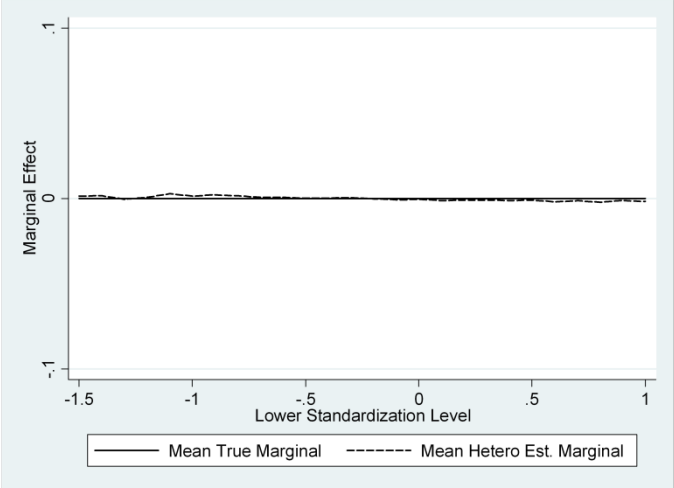


Notes: See notes to Figure A1. The only difference is as follows: Estimates are generated by Monte Carlo simulation, drawing 4,000 observations (2,000 white and 2,000 black) observations from two truncated normal distributions (one at each value of X^{I*} (in steps of $0.1 \pm 0.1 \cdot SD(X^{I*})$), *and one at each value of* $X^{I*} + .5$ (again in steps of 0.1) $\pm 0.1 \cdot SD(X^{I*})$), and estimating probit model.

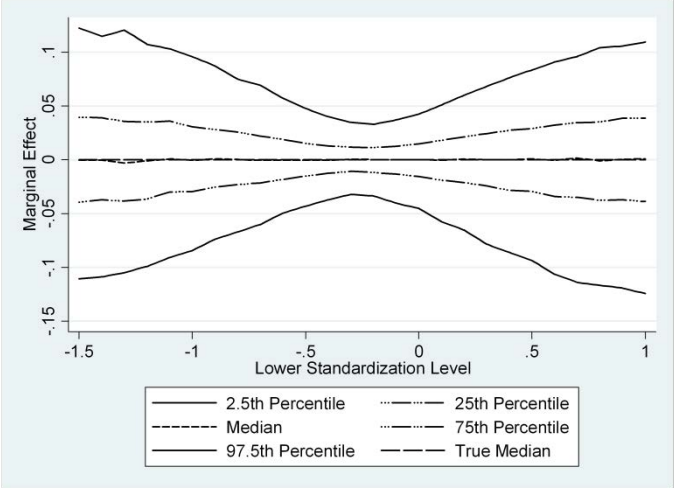
Figure A2
Monte Carlo Simulations of Probit Estimates and Marginal Effects from Simple Probit Estimation with Two Types of Applicants

No discrimination ($\gamma = 0$)

Estimates of marginal effects of black

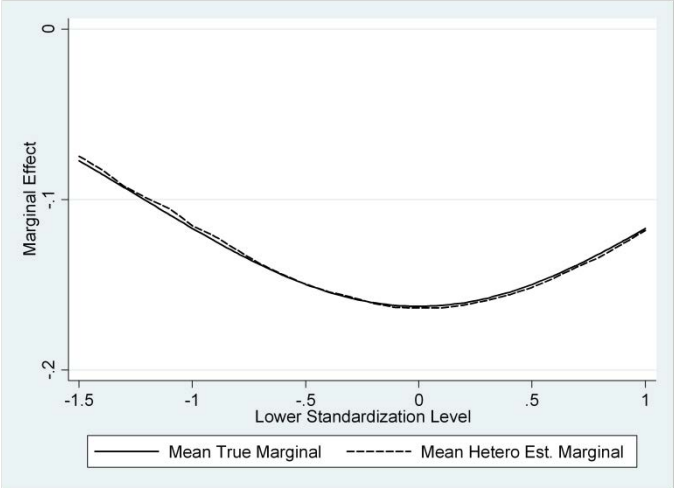


Distribution of estimates

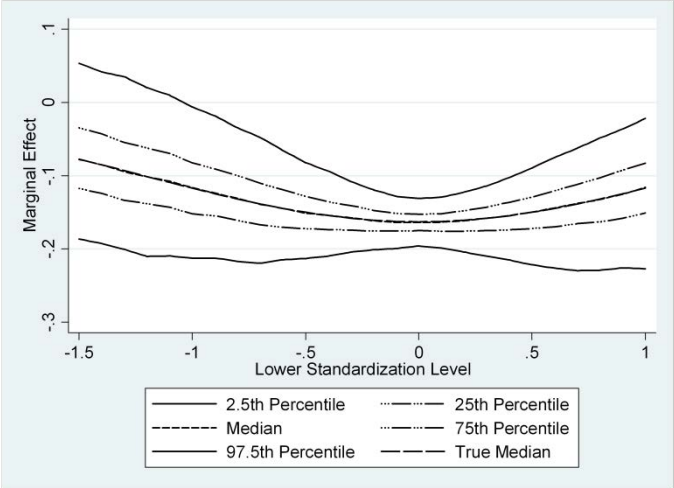


Discrimination ($\gamma = -.5$)

Estimates of marginal effects of black



Distribution of estimates

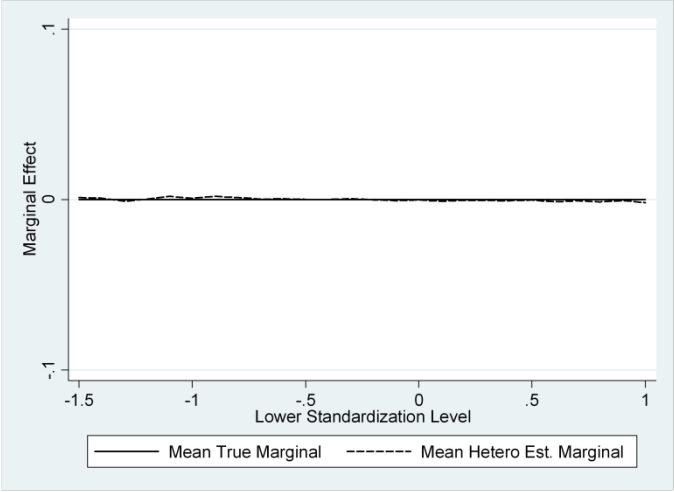


Notes: Left-hand graphs show mean true and estimated marginal effects of being black, evaluated at sample means for simulated data. Right-hand graphs show distributions of estimates. The marginal effects shown correspond to the effect of race on the latent variable, as in Equation 16'. Estimates are generated by Monte Carlo simulation, drawing 4,000 observations (2,000 white and 2,000 black) observations from two truncated normal distributions (one at each value of X^{I*} (in steps of $0.1 \pm 0.1 \cdot SD(X^{I*})$), and one at each value of $X^{I*} + .5$ (again in steps of $0.1 \pm 0.1 \cdot SD(X^{I*})$), and estimating heteroskedastic probit model. Simulation is done 5,000 times at each value of X^{I*} shown in graph. As in Figure A1, the data generating process has $X^{I*} \sim N(0,1)$, $X_B^{II} \sim N(0,1)$, $X_W^{II} \sim N(0,2.25)$, so $Var(X_W^{II})/Var(X_B^{II}) = 2.25$ (X_W^{II} and X_B^{II} are unobservable); and $\beta_I = 1$ and $c = 0$ for both blacks and whites.

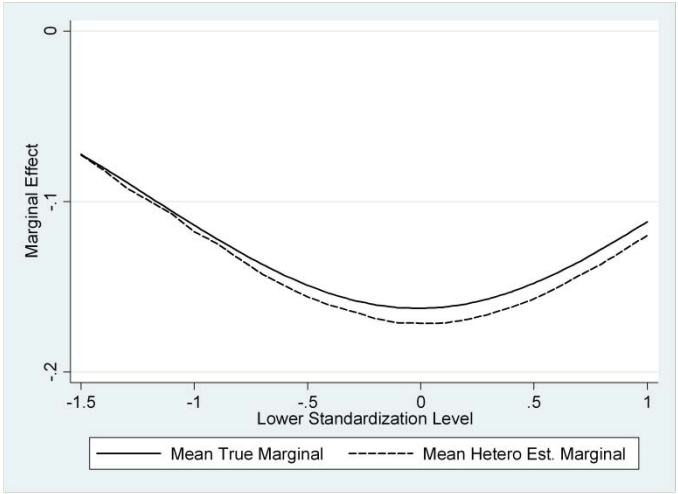
Figure A3
Monte Carlo Simulations of Heteroskedastic Probit Estimation, Estimates of Marginal Effects and Distributions

Mild violation of identifying assumption in data generating process (β_I for whites = 1.1)

No discrimination (γ = 0)

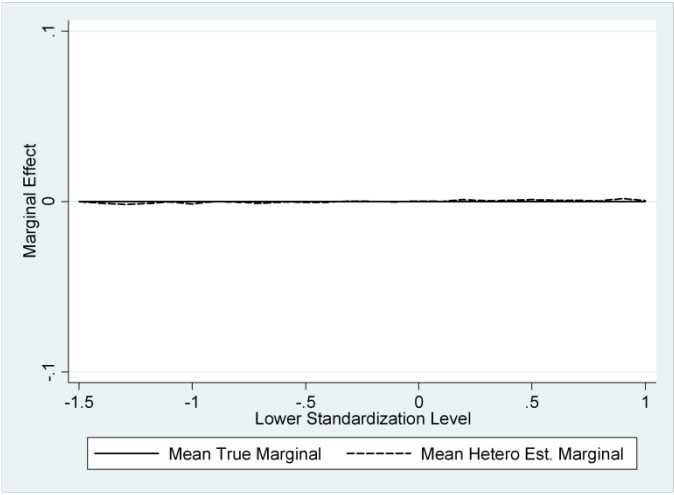


Discrimination (γ = -.5)

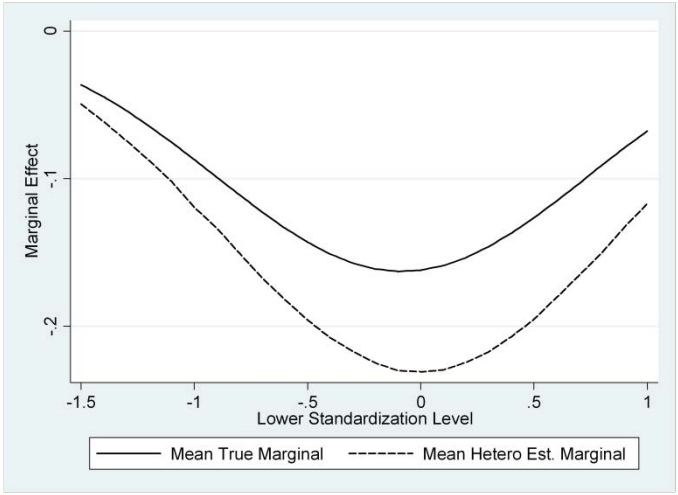


Strong violation of identifying assumption in data generating process (β_I for whites = 2)

No discrimination (γ = 0)



Discrimination (γ = -.5)



Notes: See notes to Figure A3. The only difference is that β_I is unequal for blacks and whites; it is always equal to 1 for blacks, and as indicated in the graph headings for whites.

Figure A4
Monte Carlo Simulations of Heteroskedastic Probit Estimation, with Model Misspecification Masking Higher Unobserved Variance for Whites, Estimates of Marginal Effects

Endnotes

[1] In the ensuing discussion X^I is observable and X^{II} unobservable. I put a coefficient on X^I in Equation 2 because X^I has a definable scale, whereas I treat X^{II} as scaled such that its coefficient equals one. Following HS, I assume that the coefficient on X^I is the same for blacks and whites, so discrimination is reflected only in an intercept difference. I return to this issue later.

[2] We could include resume characteristics, which should not matter for the estimate of γ' since they are randomly assigned by race. The same is true of firm fixed effects, which are orthogonal to race.

[3] As discussed below, some characteristics of workers not controlled by the researcher are observed by the employer in an audit study; but many are not.

[4] This is more problematic in correspondence studies of age discrimination, because even with many other qualifications on the resumes, if researchers give older applicants the same amount of experience as younger applicants, employers are likely to make adverse assumptions about older applicants whose resumes reflect limited work experience. See Lahey (2008) and Riach and Rich (2007) for suggestions for addressing this problem in age discrimination studies.

[5] We can treat F as statistically independent because resume characteristics are, or should be, assigned randomly (for example, Lahey and Beasley 2009). And we can always think about X^{II} as the variation in unobserved productivity that is orthogonal to X^I.

[6] Heckman (1998, footnote 7) suggests that the case with a low level of standardization and higher dispersion for whites "seems to rationalize" audit study evidence of discrimination against blacks. It is not clear, however, that we know either the level of standardization or the relative dispersion of unobserved productivity. Even though the first issue relates to observables, there is no obvious way to compare the distributions of qualifications of testers in an audit study to the population of job applicants. In fact, there are two conflicting tendencies in setting standards for

audit studies. Setting a low standard implies that call-back rates will be low, reducing the statistical power of the evidence. But setting a standard too high raises concerns about "overqualification" of candidates (for example, BM 2004, p. 995), which in an economic context presumably means that the job applicant will be expected to get better job offers, deterring employers from making offers.

[7] Conversely, the following argument can be easily modified to show that if β_I is *not* the same for blacks and whites, then γ is unidentified. The assumption that only race shifts the variance of the unobservable (Equation 13) is less consequential, since in a correspondence (or audit) study the observed characteristics of applicants other than race are essentially the same.

[8] Consistent with the earlier discussion of statistical discrimination, we might want to allow for the possibility that $E(X_B^{II}) - E(X_W^{II}) \neq 0$. In this case, we can normalize by assuming $E(X_W^{II}) = 0$ and defining $E(X_B^{II}) - E(X_W^{II}) = \mu_{BW}^{II}$. We can then replace γ in the preceding identification argument with $\gamma + \mu_{BW}^{II}$, and it is this sum of parameters, reflecting the combination of taste discrimination and the expected mean difference in the unobservable (statistical discrimination), which is identified.

[9] Some of the other controls on the resumes may be continuous, but they do not contribute to identification of the effect of race on the latent variable given that they are, by construction, orthogonal to race. By way of contrast, in audit-type studies of discrimination on a continuous outcome – such as price quotes for cars or car repairs (Goldberg 1996; Gneezy and List 2004) – the identification problem is simpler.

[10] Recall that μ in Equation 13 is normalized to zero.

[11] See Cornelißen (2005).

[12] BM actually study differences in treatment between applicants with black-sounding names and names that do not sound black. For simplicity, I discuss the results as if they capture differences between blacks and whites, which is certainly a plausible interpretation of their findings.

[13] This variable enters as a quadratic, and the effect of experience is stronger for whites up to about 16 years of experience, more than twice the mean in their sample.

[14] Also, the effect of gaps in employment is inexplicably positive, but not significant for estimates disaggregated by race.

[15] Probits estimated for males only yielded similar results for the effects of race, although the estimated coefficients of some of the productivity-related characteristics were quite imprecise or had unexpected signs. In estimating the heteroskedastic probit model for males, in some cases there were computational problems, likely reflecting these other issues regarding the estimates for males, and perhaps also the much smaller sample for males.

[16] They study sales, administrative support, clerical, and customer service jobs. The male applicants were used almost exclusively for the sales jobs, so the sex difference is identified mainly from the sales jobs.

[17] There is a small economics literature that studies discrimination based on second moments – in particular the variance of productivity – where a group with higher variance may be penalized because of risk aversion on the part of employers. The idea goes back to Aigner and Cain (1977). Dickinson and Oaxaca (2009) provide an experimental study of this type of discrimination in labor markets.

[18] One could imagine more complicated scenarios in which different firms engage in strategies that draw applicants from different parts of the distribution, in which case it is harder to

characterize how differences in the distribution of unobservables by race will impact the treatment of whites versus blacks.

[19] In principle this can be done in an audit study as well a correspondence study, although it is much harder to generate large samples in audit studies.

[20] This is the ratio of the variances of the unobservables. Note that a larger value for whites is the opposite of the common assumption in models of statistical discrimination.

[21] The true marginals are based on the heteroskedastic probit specification (Equation 16'), because the simulated data are heteroskedastic. The true marginal effect is reported as a mean because it is computed once using each simulated data set, holding the parameters fixed, and then averaged.

[22] In this case the unbiased estimate occurs at the value of −0.25 (for X^{I*}) on the horizontal axis, where the average of the upper and lower standardization levels equals zero. The reduction in bias is a little less clear in the discrimination case. To clarify, the bias in Figures A1 and A2 should be contrasted at comparable values of X^{I*}, given that Figure A2 shows the mean estimates at the lower level of standardization of X^{I*}. For example, for the discrimination case, the mean estimate of γ at $X^{I*} = 1$ in Figure A1 should be compared to the mean estimate at $X^{I*} = 0.75$ in Figure A2 (in which case this is the lower standardization level and the average is 1); the latter estimate is in fact closer to zero.

[23] In the standard marginal effect $-\psi_k\phi(S\psi)$, from Equation 15 – nearer the center of the distribution the larger estimate of ψ_k dominates the marginal effect, whereas nearer the tails the larger estimate of ψ_k lowers $\phi(S\psi)$ enough that the product $\psi_k\phi(S\psi)$ is closer to zero.

E-lit-015: Do Field Experiments on Labor and Housing Markets Overstate Discrimination? A Re-examination of the Evidence

David Neumark and Judith Rich | ILR Review 72(1): 223-252 (NBER Working Paper 22278) | 2019

https://www.nber.org/system/files/working_papers/w22278/w22278.pdf

About this source

Applying the Neumark (2012) heteroscedastic-probit correction to the 10 audit and correspondence studies with usable applicant-quality variation, the four housing studies are entirely robust - every bias-corrected estimate remains a statistically significant penalty against the minority applicant, and in Ahmed et al. the penalty on immediate showings roughly doubles (from -0.074 to -0.146). The labor-market picture is mixed: all 14 conventional estimates show discrimination, but only six survive correction. Baert et al.'s Turkish-versus-native Belgian estimate flips from significantly negative to positive and insignificant, and Lee and Khalid's Malaysian estimate flips sign outright to a significant +0.244. But the survivors include the two studies most relevant to the U.S. race-in-hiring question: Bertrand and Mullainathan's estimate gets larger in magnitude under correction (level effect -0.070 with standard error 0.040, versus -0.030 uncorrected, significant at the 10 percent level), and Carlsson and Rooth's Swedish estimate barely moves (-0.102 corrected versus -0.095), because their estimated variance ratio is 1.03. Correction is not a one-way ratchet down: Oreopoulos's Chinese and Indian estimates roughly double in magnitude, and Drydakis's sexual-orientation penalties grow from -0.384 and -0.304 to -0.476 and -0.499. In about half of all estimates the minority group has the higher estimated variance, so the bias runs toward understating discrimination as often as overstating it. The four Oreopoulos groups that lose significance mostly do so without the data rejecting the homoscedastic restriction - the authors say so explicitly, which makes those losses a precision cost of estimating a looser model rather than evidence of bias. The authors themselves suggest the robustness of Carlsson and Eriksson's housing study owes to its use of applications with substantial variation in applicant characteristics, exactly the design feature that defuses the critique.

Bound as the original PDF. Retrieved 2026-09-10. sha256 c17e8dc7de3e7ffe...

Bound into an Economic Briefing · by Claude, may contain mistakes

Full document begins on the next page.

NBER WORKING PAPER SERIES

DO FIELD EXPERIMENTS ON LABOR AND HOUSING MARKETS OVERSTATE DISCRIMINATION? A RE-EXAMINATION OF THE EVIDENCE

David Neumark
Judith Rich

Working Paper 22278
http://www.nber.org/papers/w22278

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2016, Revised August 2017

We wish to thank the following authors of studies who generously provided their raw data: Ali Ahmed, Lina Andersson, and Mats Hammarstedt; Stijn Baert, Bart Cockx, Niels Gheyle, and Cora Vandamme; Marianne Bertrand and Sendhil Mullainathan; Mariano Bosch, M. Angeles Carnero, and Lídia Farré; Magnus Carlsson and Stefan Eriksson; Dan-Olof Rooth (also with Magnus Carlsson); Nick Drydakis; Michael Ewens, Bryan Tomlin, and Liang Choon Wang; Hwok-Aun Lee and Muhammad Abdul Khalid; and Phil Oreopoulos. We thank Nick Drydakis and Philip Oreopoulos, and three anonymous referees, for helpful comments. 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.

© 2016 by David Neumark and Judith Rich. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Do Field Experiments on Labor and Housing Markets Overstate Discrimination? A Re-examination of the Evidence

David Neumark and Judith Rich

NBER Working Paper No. 22278
May 2016, Revised August 2017
JEL No. J71

ABSTRACT

There have been over 80 field experiments on traditional dimensions of discrimination in labor and housing markets since 2000, in 23 countries. These studies nearly always find evidence of discrimination against minorities. However, the estimates of discrimination in these studies can be biased if there is differential variation in the unobservable determinants of productivity or quality of majority and minority groups, so it is possible that this experimental literature as a whole overstates the evidence of discrimination. We re-assess the evidence from the 10 existing studies of discrimination that have sufficient information to correct for this bias. For the housing market studies, the estimated effect of discrimination is robust to this correction. For the labor market studies, in contrast, the evidence is less robust, as just over half of the estimates of discrimination either fall to near zero, become statistically insignificant, or change sign.

David Neumark
Department of Economics
University of California at Irvine
3151 Social Science Plaza
Irvine, CA 92697
and NBER
dneumark@uci.edu

Judith Rich
Economics and Finance Department
University of Portsmouth
Richmond Building Portland Street
Portsmouth PO1 3DE
United Kingdom
judy.rich@port.ac.uk

1. Introduction

Field experiments – specifically, audit or correspondence studies – have been used extensively to test for discrimination in markets. In audit studies of labor market discrimination, fake job candidates ("testers") of different races, ethnicities, etc., who are sometimes actors, are sent to interview for jobs (or in some early studies, apply by telephone). The candidates have similar resumes and are often trained to act, speak, and dress similarly. Correspondence studies, in contrast, use fictitious job applicants who exist on paper only (or now, electronically), and differ systematically only on group membership. The response captured in correspondence studies is a "call-back" for an interview or a closely related positive response. In contrast, the final outcome in audit studies is actual job offers. Differences in outcomes between groups are likely attributable to discrimination, although there are, naturally, some subtle issues of interpretation – including the fact that such differences can be attributable to either taste discrimination or statistical discrimination. Audit and correspondence (AC) studies have also been used to study discrimination in housing markets. In audit studies, the testers of different races, ethnicities, etc., are sent to inquire about properties for rent or sale. In correspondence studies the fictitious inquiry is submitted electronically, applying online to advertised properties for rent or sale.

AC studies are widely regarded as providing more rigorous evidence on discrimination than can be obtained from non-experimental evidence in which group membership may be correlated with unobservables.[1] Nonetheless, AC studies have come in for criticism (Heckman and Siegelman, 1993; Heckman, 1998). The most challenging criticism of these studies is that, in the standard implementation, the resulting estimate of discrimination can be biased in either

[1] The methods and empirical findings from these studies have been reviewed by Pager (2007), Riach and Rich (2002), Rich (2014), and Neumark (2016). There are, additionally, similar studies of discrimination in consumer markets (e.g., Doleac and Stein, 2013).

Does the identification critique bite - Neumark & Rich, Do Field Experiments Overstate Discrimination? (NBER WP 2017) 274

direction – or equivalently, discrimination can be unidentified. This problem arises when the variances of the unobservables differ across the groups studied, something that cannot be ruled out or easily controlled in AC studies, and indeed a difference in the variances of unobservables is at the core of early models of statistical discrimination (Aigner and Cain, 1977). This criticism – which we refer to as the "Heckman-Siegelman critique" – holds even under quite ideal conditions (detailed later) in which other potential research design flaws that Heckman and Siegelman discuss are absent.

A statistical method that can lead to unbiased estimates of discrimination using data from AC studies, relying on identifying assumption, was proposed in Neumark (2012). As explained below, most past AC studies do not have the requisite data, which are applicant or other characteristics aside from the group identifier that shift the probability of call-backs or hires. However, we have identified 10 studies of discrimination against minorities (based on race, ethnicity, or sexual orientation) in labor and housing markets conducted over the last couple of decades that do include the requisite data.[2]

These 10 studies – just like nearly all of the far greater number of AC studies that do *not* have the requisite data – find evidence of discrimination against ethnic or racial minorities, immigrants, or gays and lesbians.[3] We have obtained the original data from the authors of these studies, and our goal in this paper is to test whether this evidence is robust to confronting the data with the Heckman-Siegelman critique. Specifically, implementing the correction for bias from differences in the variances of unobservables across groups, do these studies still uniformly point

[2] The studies are: Ahmed et al. (2010); Baert et al. (2015); Bertrand and Mullainathan (2004) – the data used in Neumark (2012); Bosch et al. (2010); Carlsson and Eriksson (2014); Carlsson and Rooth (2007); Drydakis (2014); Ewens et al. (2014); Lee and Khalid (2016); and Oreopoulos (2011).
[3] For the most recent review of a large number of AC studies, see Neumark (2016).

to discrimination? Some very recent AC studies have implemented this bias correction.[4] Our goal in this paper is to revisit past studies that do not address the Heckman-Siegelman critique, to assess whether the near-uniform findings of discrimination from the large body of past research is robust to addressing this critique. We cannot re-examine all such studies. But we do, we believe, re-examine the complete set of such studies that focus on traditional dimensions of discrimination and have (accessible) the data required to address this critique.

After providing some background details on these studies, we explain the approach and report results. To summarize the results briefly, for the housing market studies the estimated effects of discrimination are robust to this correction. For the labor market studies, in contrast, the evidence is less robust; in about half of the cases covered in these studies, the estimated effect of discrimination either falls to near zero or becomes statistically insignificant, and in one the sign changes. The results for the labor market, in particular, suggest that researchers need to build into future AC studies the data and experimental design needed to address the Heckman-Siegelman critique, and that further work on different ways to eliminate bias from AC studies estimates of discrimination is warranted. More substantively, our re-examination of the evidence suggests that the overall body of experimental evidence on labor market discrimination provides a less clear signal of discrimination than one would draw from the results reported in the existing studies.

2. The field experiments covered in this paper

The field experiments re-analyzed in this paper are one of three broad types: studies of ethnic/immigrant or race discrimination in labor markets; studies of sexual orientation discrimination in labor markets; and studies of ethnic/immigrant or race discrimination in rental

[4] See Baert (2014, 2015, 2016), Carlsson et al. (2013), Neumark et al. (2015), and Nunley et al. (2015). Baert and Verhofstadt (2015) also do this, although in relation to criminal background (juvenile delinquency), which is outside the scope of discrimination studies covered in the present paper.

housing markets. Many of the details and results of these studies are discussed in Rich (2014) and Neumark (2016). Here we focus only on what is essential to understand the analysis of bias from differences in unobservables that we implement in this paper. Readers interested in more details on these specific studies, and the techniques used more generally, should see our surveys (or of course the original papers). We do not go into more detail because our goal in this paper is not to compare or critique other dimensions of these studies, but rather just to consider the robustness of the conclusions to addressing the Heckman-Siegelman critique.[5]

What distinguishes these 10 studies from the others in the literature is that they use applicants distinguished not only by race, ethnicity (including immigrant origin), or sexual orientation, but also by different levels of qualifications. In these studies, this was done to ask, in a general way, whether the evidence of discrimination by ethnicity, race, or sexual orientation differed for applicants with different levels of qualifications.[6] As discussed in the next section, however, the availability of data with variation in applicant qualifications is exactly what is needed to implement the empirical method that addresses the Heckman-Siegelman critique.

Baert et al. (2015), Bertrand and Mullainathan (2004), Carlsson and Rooth (2007), Drydakis (2014), and Lee and Khalid (2016) all used matched pairs (sets) of applicants, with two (or more)

[5] There are also field experiments investigating differences in hiring outcomes based on other characteristics, such as criminal background, mental or physical illness, facial attractiveness, veteran status, or socio-economic background or class. While these kinds of differences are not the focus of our paper (even though some could be interpreted as discrimination), the experimental designs in these papers do not generate the data needed to implement this empirical method, with the exception of Baert and Balcaen (2013), who implement this method in relation to differential treatment based on military service, and find no evidence of bias from differences in the variances of unobservables.

[6] The first study of this type (Jowell and Prescott-Clarke, 1970) considered this issue. The study compared job offer outcomes for immigrant versus white British applicants, and gave half the applications in each group higher qualifications with regard to education. (There was also variation among the immigrants only in whether they were English-speaking and whether secondary education was in Britain, although this kind of variation that does not apply equally to majority and minority groups is not as useful.) The more recent studies with such data that we re-examine in the present paper are those for which we could recover the data from authors.

applications sent to each job vacancy. Oreopoulos (2011) considered differences for many different ethnic groups (relative to native Canadians), in some cases also signaling immigrant status, and sent multiple resumes for each job vacancy. Across these studies, on the resumes used, which were either real resumes the authors found or resumes generated randomly from elements of other resumes, race or ethnicity was signalled by name, and immigrant status in addition to ethnicity was sometimes further signalled by education or work experience in a foreign country (Oreopoulos, 2011). Sexual orientation was signalled by participation in an organization active on behalf of the gay community or a gay organization.

There have been fewer studies of discrimination in housing markets in the broader literature. In the housing market experiments we re-examine, only Bosch et al. (2010) used matched pairs, while the other three (Ahmed et al., 2010; Ewens et al., 2014; Carlsson and Eriksson, 2014) sent a single rental enquiry. An accompanying message providing details on the applicant was attached, in which the researchers manipulated the information provided – ethnicity and race, as well as other qualifications or the applicant's job, which indicated ability to pay. In these studies, signaling is done by name, although Bosch et al. (2010) interpret their results for Moroccan versus Spanish names as measuring discrimination against immigrants.

Other qualifications also varied across the resumes or applications – and this variation in qualifications is essential for implementing the correction for bias from differences in variances of unobservables. The variables used in each study are described in Tables 2A, 2B, and 3, which report our results from re-analyzing the data from these studies (discussed in detail below). For example, Bertrand and Mullainathan (2004) generally sent four applications to each job. They created two matched pairs of applicants, one with low-quality background and another pair with high-quality background. The quality of the applicant varied based on labor market experience, career profiles, employment history, and skills such as employment experience gained either over

summer or while at school, volunteering, extra computer skills, certification degrees, foreign language skills, honors, or some military experience. Carlsson and Rooth (2007) signalled similar additional information on applicants as Bertrand and Mullainathan, as well as different spells of unemployment, work experience over the summer, overqualified or not, personality traits, and cultural and sporting activities listed as hobbies and interests. Oreopoulos (2011) varied the information provided on the extent of foreign education and foreign experience as well as language skills and certification and masters degrees. Drydakis (2014) used an accompanying cover letter to provide more favorable information about applicants in some cases, including mentioning grades, previous job responsibilities and tasks, and personality characteristics associated with work commitment; these same applicants also included letters of references that more strongly signalled positive work traits such as teamwork and loyalty to the firm. Lee and Khalid (2016) varied factors such as private versus public university, grades, and English proficiency.

In the housing market tests, researchers manipulated the information on the applicant, using an accompanying message, to explore the impact of basic, negative, or positive information – such as habits (smoking, exercise, and nightclub attendance, in Carlsson and Eriksson, 2007), variation in smoking and credit rating (in Ewens et al., 2014), and information on positive characteristics like work history, education, lack of payment complaints, etc. (Ahmed et al., 2010) or stable occupations and contracts (Bosch et al., 2010).

The richness and number of qualifications that researchers chose to vary across the applicants differ quite a bit across these studies. For the labor market studies, these qualifications generally pertain to education, experience, and skills, but sometimes extend to attempts to convey something about the applicant's personality or hobbies, the order of the application, and other things. One of the housing studies (Carlsson and Eriksson, 2014) tries to provide information on the applicant's

lifestyle, which could be relevant to a potential landlord. We do not discuss the different qualifications used in each study in detail, but list them for each study in the tables reporting the statistical analysis (Tables 2A and 2B for the labor market studies, and Table 3 for the housing market studies). The reader will note that we also list other features of the ads that could affect the probability of a call-back – such as characteristics of the job or the apartment. We include these because – as explained in the next section – the statistical method is informed by differences in the coefficients between the two groups studied in *any* of the factors that can affect call-backs.

3. Findings from the field experiments covered in this paper

Table 1 summarizes the conventional results from the 10 studies we re-examine, as well as giving basic information about them, including the years covered, the groups covered, and the outcomes. The original studies report results in different ways, varying between chi-square/Fisher exact tests, binomial tests, or tests of the null hypothesis that there is no difference in the call-back rate between the groups, typically controlling for other aspects of the resumes. However, here we report results on a consistent basis for all studies – marginal effects from probit models using the full set of resume characteristics included in the data – which we have estimated from data provided by the authors of these studies.[7]

As reported in Table 1, the six labor market experiments covered in Panel A all find statistically significant evidence of discrimination against either ethnic minorities, blacks, or gays and lesbians. The estimated differentials by racial and ethnic groups are in the same range – an approximately 0.03 to 0.15 lower probability of a call-back. These are on somewhat different

[7] Details on the control variables, the standard errors, etc., are provided in tables discussed below. Not surprisingly, the results in Table 1 closely parallel the conclusions of the original papers – however they report their results – although they are not always identical.

baseline rates of call-backs, but the call-back rates also do not vary that much across these studies.[8] The two estimates from Drydakis (2014), for discrimination against gays and lesbians in Cyprus, are much larger (although the baseline call-back rates are much higher too).

The four housing market studies similarly find consistent evidence of discrimination against minorities. The range of estimates is fairly tight (a 0.09 to 0.17 lower call-back rate). Thus, every one of these studies points to evidence of discrimination against the minority group.

The conclusions from these studies strongly echo the broader literature, in which nearly every study finds evidence of discrimination in labor or housing market on the basis of race or ethnicity (Rich, 2014; Zschirnt and Ruedin, 2015; Neumark, 2016; Quillian et al., n.d.), as do the smaller number of studies of discrimination based on sexual orientation (Neumark, 2016). The question this paper addresses is whether this near-uniform evidence of discrimination from field experiments is an accurate reflection of discriminatory behavior, supporting a conclusion that discrimination really is this consistent and pervasive, or whether the evidence in at least some of these studies might reflect biases stemming from differences in the variance of unobservables across groups – the problem highlighted by the Heckman-Siegelman critique.

Some of the studies also include female and male applicants, or more broadly test for discrimination along multiple dimensions, including sex and age (Carlsson and Eriksson, 2014). We do not focus, in this paper, on evidence on discrimination based on sex or age. The broader literature focuses far more on race and ethnicity (and more recently on sexual orientation), and – as we have noted – delivers a near-uniform finding of discrimination against minorities. The evidence of sex discrimination is less robust, and tends to point less to discrimination against women, and

[8] One might wonder about apparent evidence of discrimination against British immigrants in Canada; indeed, we will see in implementing the correction for the Heckman-Siegelman critique below that this evidence appears to be spurious.

more to the importance of sex norms for jobs in whether male or female applicants received more call-backs (Neumark, 2016). And recent evidence from a large-scale correspondence study of age discrimination yields ambiguous results for men, but not women (Neumark et al., 2016).

We next provide a brief discussion of the approach used to correct for the bias in estimates of discrimination from the standard field experiment design, and then present our re-examination of the data from the 10 studies we have identified that have the requisite data to implement the method in Neumark (2012) to correct the estimates for bias from differences in the variances of unobservables.

4. Addressing the Heckman-Siegelman critique

There are quite a few critiques of AC studies aside from the one we focus on here. Most of them are laid out in Heckman and Siegelman (1993), and discussed further in Neumark (2012) in the context of the framework laid out in this section. Some of the more important critiques – such as the possibility of "experimenter effects," and small differences between applicants that can matter a lot when applicants are matched on so many characteristics – can be addressed by using correspondence studies instead of audit studies, and indeed most recent research uses the correspondence study technique. The Heckman-Siegelman critique is of particular importance because it applies equally well to correspondence studies, even under otherwise ideal conditions such as no *mean* differences in unobservables between groups, but only differences in the *variances* of unobservables. And this critique is salient because nothing in the research design rules out differences in the variances of unobservables, and indeed – as noted earlier – these differences are foundational in models of statistical discrimination. We first lay out a basic framework for the analysis of data from an audit or correspondence study, and then explain the

bias and the correction.[9]

Non-experimental regression-based approaches testing for and measuring discrimination use data on the groups in question in a population, introducing regression controls to try to remove the influence of group differences in the population that can affect outcomes (Altonji and Blank, 1999). Correspondence (and audit) studies, in contrast, create an artificial pool of labor market participants among whom there are supposed to be no average differences by group. This is clearly a potentially powerful strategy, because if we have, e.g., a sample of blacks and whites who are identical *on average*, because race is randomly assigned to a subset of similar resumes, then in a regression of the form

$$Y = \alpha + \beta B + \varepsilon \,, \qquad (1)$$

where Y is the outcome and B is a dummy variable for blacks, ε is uncorrelated with B, so that the OLS estimate $\hat{\beta}$ (or simply the mean difference in Y) provides an estimate of the effect of race discrimination on Y.[10]

Of course, most of the earlier regression studies focus on wages, whereas AC studies focus on hiring. If an employer is free to pay a lower wage to blacks, for example, then in the context of the Becker employer discrimination model, why discriminate in hiring? One common interpretation is that there is an equal wage constraint – perhaps due to a minimum wage, or because anti-discrimination laws are more effective at rooting out wage discrimination than hiring discrimination. Alternatively, in the simple model, employers with stronger discriminatory tastes than the marginal employer will discriminate in hiring. As we make clear below, however, this framework does not only detect taste discrimination à la Becker.

---

[9] This section draws heavily on Neumark (2012), while avoiding many details that a reader can find in that paper.

[10] For simplicity, the discussion here is couched solely in terms of blacks and whites.

To provide a more formal framework, suppose that productivity depends on two individual characteristics (standing in for a larger set of relevant characteristics), $X' = (X^I, X^{II})$, so that productivity is $P(X')$. X^I is what the firm observes, and X^{II} is unobserved by firms. It is simplest, for now, to think of Y as continuous, such as the wage offered, although in fact in AC studies we should think of it as latent productivity leading to a decision to hire/call-back or not.

Define discrimination as

$$Y(P(X'), B=1) \neq Y(P(X'), B=0) . \quad (2)$$

Assume that $P(.,.)$ is additive, so

$$P(X') = \beta_L X^I + X^{II}, \quad (3)$$

where the coefficient of X^{II} is normalized to one as it is unobservable, and

$$Y(P(X'), B) = P + \gamma B. \quad (4)$$

Discrimination against blacks implies that $\gamma < 0$, so that blacks are paid less than or perceived as less productive than whites who are actually equally productive.

In correspondence studies, researchers create resumes that standardize the productivity of applicants at some level. Denote expected productivity for blacks and whites, based on what the firm observes, as P_B^* and P_W^*. Y is observed for each tester, so each test – the outcome of applications to a firm by one black and one white tester/applicant – yields an observation

$$Y(P_B^*, B = 1) - \mathrm{Y}(P_W^*, B = 0) = P_B^* + \gamma - P_W^* . \quad (5)$$

Given that the correspondence study design sets $P_B^* = P_W^*$, we should be able to estimate γ easily from these data, by simply running a regression of Y on the dummy variable B and a constant. (Some potential complications are discussed in Neumark, 2012).

A correspondence study can preclude systematic differences between groups in observables and experimenter effects. But there can still be assumed differences in means between

groups despite the groups using matched resumes. In equation (5) above, $P_B^* = E(\beta_1 X_B^I + X_B^{II} | X_B^I, B = 1)$, and similarly for P_W^*. Assuming randomization, and with $X_B^I = X_W^I = X^I$, the right-hand side of equation (5) reduces to $\gamma + E(X_B^{II} | X^I, B = 1) - E(X_W^{II} | X^I, B = 0)$, implying that we only identify γ if $E(X_B^{II} | X^I, B = 1) = E(X_W^{II} | X^I, B = 0)$. Employers may have different expectations about the mean of X^{II} for blacks and whites, conditional on what they observe, which a labor economist would label statistical discrimination. Although economists are interested in distinguishing between statistical and taste discrimination, both are illegal under U.S. law and both also appear to be illegal under European Union law.[11] Moreover, it is challenging to distinguish between the two models. Thus, this issue is put aside, and the discrimination estimates from the studies considered in this paper interpreted as the sums of taste and statistical discrimination.[12]

That is not to suggest that researchers using AC methods have not tried to distinguish between taste and statistical discrimination. The idea exploited in most studies is that when the applications include a richer set of applicant characteristics, it is less likely that statistical discrimination plays much of a role in group differences in outcomes (e.g., Ewens et al., 2014).

---

[11] As discussed in Neumark (2016), the U.S. Code of Federal Regulations (29, § 1604.2) defines as illegal discrimination "The refusal to hire an individual because of the preferences of coworkers, the employer, clients or customers …" But it also states "The principle of nondiscrimination requires that individuals be considered on the basis of individual capacities and not on the basis of any characteristics generally attributed to the group. There is not as explicit a prohibition of statistical discrimination in the European Union (EU). Article 2 of the EU's Directive 2000/43/EC prohibits both "direct" and "indirect" discrimination, but these appear to line up, respectively, with disparate treatment and disparate impact in the U.S. context (see http://eur-lex.europa.eu/legal-content/en/ALL/?uri=CELEX%3A32000L0043, viewed December 2, 2015). However, other material suggests that statistical discrimination is covered by direct discrimination (OECD, 2013, p. 195).

[12] Indeed, it seems that we could also include implicit discrimination (e.g., Bertrand et al., 2005). Implicit discrimination posits a different reason for undervaluing the productivity of a group of workers, which can lead to different policy levers to combat it. But if it arises when employers evaluate applicants in AC studies, the empirical implication for the framework developed here would likely be the same as the implication of taste discrimination.

Effectively, one tries to eliminate the term $E(X_B^{II}|X^I, B = 1) - E(X_W^{II}|X^I, B = 0)$ from the estimated difference in hiring rates to see how much of the overall difference in hiring rates is accounted for by this difference in expectations, which corresponds to statistical discrimination.[13]

Oreopoulos (2011) and Ewens et al. present perhaps the most thorough attempts at discerning between these hypotheses about discrimination in AC studies. Oreopoulos uses the approach of adding information (e.g., on country of education, to signal English language skills) to see whether estimated hiring gaps fall, as well as examining differences in hiring gaps for occupations across which the importance of statistical discrimination likely varies. In many cases, he does not find evidence consistent with statistical discrimination, despite evidence from a survey of participating employers that they used name, or country of education or experience, as a signal of potential language problems.

Ewens et al. (2014) specifically allow for the mean and variances of unobservables to differ across groups (as in Aigner and Cain, 1977), and examine whether the differential treatment by race is more consistent with statistical discrimination (both first- and second-moment) or taste-based discrimination. Although they do not correct for differences in variances of unobservables, they demonstrate that group differences in outcomes may decrease when more information is provided and they argue that the evidence is consistent with statistical discrimination. In particular, they demonstrate that the differences in outcomes across groups vary with the differences in racial composition across neighborhoods in a way that is consistent with the hypothesized differences in variances of unobservables across groups.

One could presumably use the method described below for resumes with varying amounts of information, to recover unbiased estimates under different information treatments and hence try

[13] Neumark (2016) provides many examples, and also some criticisms of this approach.

to gauge the relative importance of taste and statistical discrimination. However, this issue is not the focus of our analysis in this paper. Instead, our focus is re-examining the 10 studies identified earlier and investigating whether the uniform evidence of discrimination from these studies persists once account is taken of the Heckman-Siegelman critique.

The issue raised by the Heckman-Siegelman critique arises from the potential for differences across groups in the variances of the unobservables – which is equally problematic even in the ideal condition of no assumed mean difference. To see how the difference in variances can drive differences in the results of the analysis of data from an AC study, it is most natural to think of equation (1) as a latent variable model for productivity, with applicants having to exceed some productivity threshold with sufficiently high probability (where α in equation (1) can also include observables that vary across individuals that affect productivity, which we have denoted X^I).

To isolate the problem, consider the best-case scenario where $E(X_B^{II}|X^I, B = 1) = E(X_W^{II}|X^I, B = 0)$ – i.e., there is no statistical discrimination regarding levels. But the standard deviations of the unobservables, denoted σ_B^{II} and σ_W^{II}, need not be equal.[14]

Assume the applicant is called back (hired) if there is a sufficiently high probability that their productivity exceeds a given threshold. In this case, the inequality $\sigma_B^{II} \neq \sigma_W^{II}$ combined with the design of AC studies results in a biased estimate of discrimination; worse, we cannot necessarily even sign the bias.

To see the intuition, recall that the key feature of the usual design of AC studies is using similar resumes on the applicants in different groups. This requires choosing a particular level of

[14] Neumark assumes homoscedasticity within groups, and thus suppresses conditioning on X_B^I and X_W^I.

the quality of the resumes. Suppose, for example, that the research design standardizes X^I at a low level, denoted X^{I*}. Employers care about how likely it is that the sum $\beta_L X^I + X^{II}$ exceeds some threshold. Given the low value X^{I*}, this is more likely for a group with a high variance of X^{II}. Thus, even in the case of no discrimination ($\gamma = 0$), the employer will favor the high-variance group. Conversely, if standardization is at a high level of X^{I*}, the employer will favor the low-variance group. Because researchers do not have information on the population of real applicants to the jobs studied, there is no definitive way to know whether X^{I*} is high or low relative to the actual distribution, and hence no way to sign the bias. As discussed in more detail below, note that the variances of unobservables affect which group gets more call-backs only because of the research design standardizing the resumes at a particular level (when the level of standardization is not at the central tendency of the distribution).

The technique developed in Neumark (2012) to correct for the bias from differences in the variances of unobservable characteristics relies on the experimental study having extra information that explores the impact of different productivity or quality characteristics (creating applicants who have different levels of qualifications, for example). As long as some of these characteristics have the same effects in the latent variable model for the probability of an offer – the key identifying assumption – this extra information allows the effect of the difference in variances between the groups' unobserved characteristics on the responses to be isolated from the role of discrimination in evaluating applicants. That is, it allows separate identification of the relative variances in the unobservables and the discrimination coefficient, γ.[15]

It is rare that correspondence studies include variables that shift the call-back probability,

[15] To reiterate, for the purposes of simplification, it is assumed $E(X_B^{II}|X^I, B = 1) = E(X_W^{II}|X^I, B = 0)$. Without this assumption, references to γ in the remainder of this section should be read as references to $\gamma + E(X_B^{II}|X^I, B = 1) = E(X_W^{II}|X^I, B = 0)$ – i.e., the sum of taste and statistical discrimination.

because these studies typically create one "type" of applicant for which there is only random variation in characteristics that are not intended to affect outcomes. However, the 10 studies discussed in Section 2 have this information – as in Bertrand and Mullainathan (2004), whose data Neumark (2012) used to illustrate this method for correcting for the bias in AC studies. Applying this method to the studies re-examined in this paper therefore allows us to determine whether the measures of discrimination from conventional analyses of the data in these studies provided unbiased estimates of discrimination, or instead either overstated or understated discrimination.[16]

The intuition behind the solution stems from the fact that a higher variance for one group (say, whites) implies a smaller effect of observed characteristics on the probability that a white applicant meets the standard for hiring. Thus, information from a correspondence study on how variation in observable qualifications is related to call-backs can be informative about the relative variance of the unobservables, and this, in turn, can identify the effect of discrimination. Based on this idea, the identification problem identified by the Heckman-Siegelman critique is solved by invoking an identifying assumption – specifically, that the effect of applicant characteristics that affect perceived productivity and hence call-backs have equal effects across groups – along with the testable requirement that some applicant characteristics affect the call-back probability (since if all the effects are zero we cannot learn about $\sigma_B^{II}/\sigma_W^{II}$ from these coefficient estimates).

In a probit specification, for example, we know that we can only identify the coefficients of the latent variable model for productivity relative to the standard deviation of the unobservable. In this case, we effectively have two probit models, one for blacks and one for whites. If we normalize σ_W^{II} to one, then for a characteristic (Z) that affects the call-back rate, we identify its

[16] For recent code to implement the estimator, we direct readers to the code used in Neumark et al. (2016), on the website of the American Economic Review (click on "Data Set" on the webpage at https://www.aeaweb.org/articles?id=10.1257/aer.p20161008).

coefficient (δ_W) relative to σ_W^{II}, or δ_W/σ_W^{II}. However, if we assume that $\delta_W = \delta_B$, then we do not need to impose the normalization that $\sigma_B^{II} = 1$, but instead can identify $\sigma_B^{II}/\sigma_W^{II}$ from the ratio of the coefficients on Z in the probit for whites versus blacks, which in turn allows us to identify γ. The estimation can be done using a heteroscedastic probit model. Finally, when there are *multiple* productivity-related characteristics that shift the call-back probability Z_k (k =1,…, K), there is an overidentification test because the ratio of coefficients on each Z, for whites relative to blacks, should equal $\sigma_B^{II}/\sigma_W^{II}$.[17]

The heteroscedastic probit model estimates can be decomposed into the estimated differential due to differences in γ, and the estimated differential due to differences in the variance of the unobservables. In generic notation, let the latent variable depend on a vector of variables S and coefficients ψ, and the variance depend on a vector of variables T, which includes S, with coefficients θ. The elements of S are indexed by k. For a standard probit model, coefficient estimates are translated into estimates of the marginal effects of a continuous variable S using

$$\partial P(\text{call-back})/\partial S_k = \psi_k \phi(S\psi) \qquad (6)$$

where S_k is the variable of interest with coefficient ψ_k, $\phi(.)$ is the standard normal density, and the standard deviation of the unobservable is normalized to one. Typically, this is evaluated at the means of S. When S_k is a dummy variable such as race, the difference in the cumulative normal distribution functions is often used instead, although the difference is usually trivial.

The marginal effect is more complicated in the case of the heteroscedastic probit model, because if the variance of the unobservable differs by race, then when race "changes" both the

[17] Indeed, the identifying restriction $\delta_W = \delta_B$ only has to hold for subsets of the characteristics that shift the call-back probability, and one can rely only on this subset if the overidentification test for a larger set of resume characteristics fails (see Neumark, 2012).

variance and the level of the latent variable that determines hiring can shift. As long as we use the continuous version of the partial derivative to compute marginal effects from the heteroscedastic probit model, there is a unique decomposition of the effect of a change in a variable S_k that also appears in T into these two components. In particular, denoting the variance of the unobservable $[exp(T\theta)]^2$, with the variables in T arranged such that the k^{th} element of T is S_k, then the overall partial derivative of P(call-back) with respect to S_k is

$$\partial P/\partial S_k = \phi(S\psi/exp(T\theta)) \cdot \{\psi_k/exp(T\theta)\} + \phi(S\psi/exp(T\theta)) \cdot \{(-S\psi \cdot \theta_k)/exp(T\theta)\}.^{18} \qquad (7)$$

The first part of the sum in equation (7) is the partial derivative with respect to changes in S_k affecting only the level of the latent variable – corresponding to the counterfactual of S_k changing the valuation of the worker without changing the variance of the unobservable. The second part is the partial derivative with respect to changes via the variance of the unobservable. In the analysis below, these two separate effects are reported as well as the overall marginal effect, and standard errors are calculated using the delta method.[19]

This discussion raises the issue of what we are trying to measure in audit and correspondence studies. Focusing on γ, the structural effect of race, captures the potential discounting by employers of black workers' productivity à la Becker (and possibly statistical discrimination about the mean of X^{II}). But as shown, employers could treat blacks and whites differently in hiring because of different variances of the unobservable. If the latter is accepted as

[18] See Cornelißen (2005).

[19] Because the formula for the derivative based on a continuous variable yields this unique decomposition, it is used below – and also to interpret the simple probit estimates, as in Table 1. The implied partial derivatives from the probit using the formula for a discrete variable (or computing the partial derivative for each sample observation and averaging, as is now more standard) were very similar. One can decompose the partial derivative from the heteroskedastic probit model based on the partial derivative for discrete variables calculated from difference in the cumulative normal distribution functions, but then the decomposition is not unique.

a meaningful measure of discrimination, we might not want to eliminate it.

There are two reasons why the coefficient γ is the focus of interest. First, to the best of our knowledge, differential treatment based on assumptions (true or not) about variances have not been viewed as discriminatory in the legal literature. Second, and probably more important, the taste discrimination (and possibly "first-moment" statistical discrimination) that correspondence studies capture in γ generalizes from the correspondence study to the real economy. In contrast, the difference in treatment based on differences in the variances of unobservables is an artifact of the design of correspondence (or audit) studies – in particular, the standardization of applicants to particular, and similar, values of the observables, relative to the actual distribution of observables among real applicants. If, instead, a study used applicants that replicated the actual distribution of applicants to the employers in the study, there would be no bias – in the setting described here – from different variances of the unobservables. This is discussed in detail in Neumark (2012).

That is not to say, however, that there cannot be discrimination based on second moments with, for example, risk averse firms. In that sense, one can potentially interpret the bias correction and decomposition not as separating out real versus spurious discrimination, but rather first-moment versus second-moment discrimination. We could imagine, for example, that risk-averse firms are less likely to call back (or hire) workers with more uncertain productivity, even when on average they are as productive as another group. However, the potential difficulty with this interpretation is that we do not uniformly find that the minority group that experiences discrimination according to the conventional analysis generally has a higher variance of the unobservable; indeed, in both the labor market studies and the housing market studies we analyze, this is the case in just about half of the estimates. This is a further reason why, in the remainder of the paper, we interpret the evidence as isolating discrimination by adjusting for differences in the variances of unobservables.

5. Results from re-examination of field experiments with quality variation across resumes

Labor market field experiments

We report the results for the re-analysis of the datasets from the labor market field experiments in Tables 2A and 2B. Turning to the first set of labor market studies covered in Table 2A, we first report the estimated discrimination coefficient (γ, in the equations from above) in the first row of the table (Panel A). These match the estimates in the last column of Table 1, and have already been summarized.

Panel B turns to the heteroscedastic probit estimates that correct for biases from differences in the variance of unobservables. The "Controls" entry toward the bottom of the table lists the resume characteristics including those likely to shift the call-back rate (like education, skills, etc.).[20] The first row of Panel B reports the overall effect from the heteroscedastic probit estimates. These are similar to the probit estimates. The next two rows of the table report the key results from the decomposition of the heteroscedastic probit estimates. The "level" effect (labelled "Marginal effect through level (unbiased)" in the table) is the unbiased estimate, and the "variance" effect reflects the bias from correspondence study design, arising because of the interaction between the quality of the resumes sent out (relative to the actual distribution) and differences in the variances of unobservables.

Looking at these estimates, for the first study – the Baert et al. (2015) experiment on discrimination against Turkish job applicants relative to natives in Belgium – the evidence of

[20] Some studies include resume characteristics that are not independent of minority group status. For example, Oreopoulos (2011) indicates, for some of his ethnic groups, that some education or experience occurred in a foreign country. This is useful for asking what might explain variation in the amount of discrimination immigrants face, which is the focus of his study. But it does not fit into the narrower question considered in this paper of discrimination against the minority group per se. Hence, we only use resume characteristics that are constructed to be orthogonal to minority group status.

discrimination completely disappears in the heteroscedastic probit estimates. In both columns (1) and (2) – the first for a call-back, and the second for an immediate interview – the negative and significant coefficient estimate on the indicator for Turkish applicants becomes positive and statistically insignificant.

In contrast, the estimated effect through the variance is negative and significant, implying that the study design generates bias towards finding evidence of discrimination. The next row of the table reports that the ratio of the estimated standard deviations of the unobservables for minority versus non-minority candidates is around 0.5, indicating a lower variance of unobservables for the Turkish applicants. In terms of the model, the reduction in estimated discrimination coupled with a lower variance of unobservables for minorities implies that on average the resumes in this study were of relatively low quality compared to what employers see; thus, the low variance group is less likely to be of sufficiently high quality on the unobservables to merit a call-back, and the difference in variance creates a bias towards finding discrimination against Turkish applicants.

Below the decomposition estimates, the table reports some additional diagnostic test results. First, it reports the p-value from the overidentification test that the ratios of the skill coefficients between (in this case) Turkish and native applicants are equal across all of the skills/resume characteristics. The p-value is 0.97 in column (1) and 0.93 in column (2), indicating that we do not reject the overidentifying restrictions. On the other hand, in this case, as reported in the next row, the data tend to reject the restriction to the homoscedastic specification; the p-value from a likelihood ratio test is 0.01 in column (1) and 0.10 in column (2). The final test result reported is whether the ratio of variances of the unobservables equals one; this is rejected strongly in both columns (a result we expect would to parallel to some extent the likelihood ratio test).

Thus, for the Baert et al. study, application of this method of correcting for bias from

differences in the variances of unobservables very much overturns the evidence of ethnic discrimination. There is one additional point to make with reference to the more general earlier discussion about interpreting the effect through the variance. One might refer to the negative (and significant) estimates on "Marginal effect through variance" as suggesting that the evidence of discrimination has not gone away, but simply been "displaced" to show up in the variance. We have already explained why, in the context of the method and underlying model used in this paper, the estimated effect through the variance is an artifact of the study, and would not be expected to be replicated in the real world. Similarly, it would not be replicated if the study had used high-quality resumes, or a distribution of resumes that matched the distribution employers actually see. An alternative hypothesis, though, is that the effect of variance is real, and reflects employer risk aversion rather than how the employer evaluates the likelihood that an applicant exceeds a call-back/hiring threshold, given the resume. However, if there is risk aversion, then high-variance groups would be penalized. That is inconsistent with the evidence from the Baert et al. data, since the minority applicants are estimated to have lower variance.[21]

Having gone through the results for the first study in detail, the results for the other labor market studies can be covered more succinctly. The Carlsson and Rooth (2007) study of discrimination against Middle Easterners in Sweden asks a very similar question to Baert et al. (2015). In this case, however, the conclusions are scarcely affected by addressing the Heckman-Siegelman critique. The estimated marginal effect through the level (−0.102) is very similar to the simple heteroscedastic probit estimate (−0.095), and the estimated marginal effect through the

[21] This may be too strong a statement, since if employers actually evaluate applicants based on their assumed variance of the unobservable, the statistical model might be different. We are not aware of any field experiments that have tried to incorporate risk aversion, although this might be fruitful. Dickinson and Oaxaca (2009) provide a lab experiment study of this type of discrimination in labor markets.

variance is close to zero (0.007) and estimated precisely. In this case the ratio of the estimated standard deviations of the unobservable for minorities relative to non-minorities is very close to one (1.03), which implies – in terms of the Heckman-Siegelman critique – that there is unlikely to be any bias regardless of the quality of the artificial resumes relative to the population of resumes that the employer sees, which is consistent with the robustness of the evidence for this study. Note also that the data do not reject the overidentifying restrictions, nor do they reject the restriction to the homoscedastic model or that the ratio of standard deviations equals one – not surprising given the estimates.

The Drydakis (2014) study looks at discrimination against gays and lesbians. In this study, also, correcting for potential bias from differences in the variances of the unobservables does not alter the conclusion much. Indeed, the estimated effect of being gay or lesbian is larger negative (−0.476 or −0.499) after correcting for this bias, relative to the overall effect of −0.384 for gays and −.304 for lesbians. For both groups, the estimated variance of the unobservable is quite a bit larger than for straight men or women, with a ratio of standard deviations of 1.59 for gay versus straight men, and 2.27 for lesbian versus straight women. The combination of a higher variance for gays or lesbians with a larger estimate of discrimination would imply that the resumes were of low quality relative to the distribution, which would lead employers to favor the high variance group and generate a bias towards zero in the estimate of discrimination.

Note that for the Drydakis analyses there is strong evidence against the homoscedastic probit model and marginally significant evidence against equal standard deviations. Also, for the analysis of gay versus straight men the overidentifying restrictions are rejected at the 10-percent level. This last result prompted us to estimate a less restrictive model that did not restrict the effects of two of the resume characteristics to be the same across gay and straight men – chosen based on the estimates indicating that these interactions did not fit the expected pattern if the

coefficients in the latent variable model were equal and only the variances of the unobservables varied.[22] In this case the overidentification restrictions were no longer rejected (the p-value was 0.751), yet the estimates were very similar to those reported in column (5) of Table 2A.

Lee and Khalid (2016) study discrimination against Malays (versus Chinese), in the private sector in Malaysia.[23] In this case, the conclusions are dramatically affected by addressing the Heckman-Siegelman critique, as the estimated marginal effect through the level changes sign and becomes significant and positive – consistent with discrimination in favor of Malays.[24] In contrast, the estimated marginal effect through the variance is large, negative, and significant (−0.445). In this case the ratio of the estimated standard deviations of the unobservable for Malays relative to Chinese is very low (0.11). The combination of a lower variance for Malays with a smaller (indeed, opposite-signed) estimate of discrimination would imply that the resumes were of low quality relative to the distribution for jobs included in the study, which would lead employers to favor the high variance group and generate a bias towards discrimination in favor of Chinese applicants. Note also that the data do not reject the overidentifying restrictions.

---

[22] These were the indicators for a high-quality resume (more experience) and for resume type. These were chosen because the estimated signs of the interactions relative to the signs of the main effects were rather strongly inconsistent with what would be predicted based on the higher estimated variance of the unobservable for gays. Note that the model is identified as long as the effects of *some* variables that shift the call-back probability are restricted to be equal across the two groups; this restriction does not have to hold for all of them, and can be relaxed by adding interactions between the group indicator and the resume characteristic to the heteroscedastic probit model.

[23] Malays are not the minority group, although we retain that label in the table to be consistent with other studies. Lee and Khalid (2016) discuss issues related to potential discrimination against Malays in the private sector, including affirmative action for Malays in public education that may lead Malay graduates to be less preferred. Their sample size with controls is a bit smaller than ours (see their Table 4), because they also include data on the companies in the study; these data are not always available, and the company data were not provided to us.

[24] While this change in results is striking, there are also findings in the Lee and Khalid paper that do not cleanly fit the expected story of discrimination against Malays. In particular, they find stronger anti-Malay discrimination in hiring for private university graduates, where affirmative action in education is *not* implemented.

Turning to the remaining labor market studies, in Table 2B, Oreopoulos (2011) studies outcomes for six immigrant groups relative to native Canadians. It turns out that for two of these groups – Chinese and Indian – the evidence of discrimination remains significant after addressing the Heckman-Siegelman critique, and is actually stronger, with estimates changing from around −0.05 to −0.10 or greater. For both groups, the estimated variance of the unobservable is larger for immigrants than for natives, which appears to interact with the applicants being low quality so that the higher variance biases the estimate of discrimination from the standard probit towards zero. In contrast, for the other four groups – Chinese-Canadian,[25] Pakistani, Greek, and British – there is no longer significant evidence of discrimination. Note that in two cases – Pakistani and Greek – the point estimate of the marginal effect of minority group membership through the level is still a large negative number, but is insignificant. In contrast, for the British, the point estimate is no longer negative.

Turning to the other diagnostics, in every case for the Oreopoulos analysis, the overidentification restrictions are not rejected. Similarly, with the exceptions of the analysis for the Chinese applicants, the data do not reject the restriction to the homoscedastic model. Thus, in this case we are sometimes failing to find evidence of discrimination because we are estimating a more flexible model even when the data do not reject a more restrictive model that provides evidence of discrimination – and the results for the Pakistani and Greek applicants are notable in this regard. This poses the usual trade-off of bias versus precision, although generally speaking labor economists are willing to estimate less restrictive models that eliminate bias at the risk of decreased precision. Regardless, it seems reasonable to conclude that the re-analysis of the Oreopoulos data indicates far less robust evidence of discrimination than the original study.

[25] This refers to an English first name and a Chinese last name.

Finally, column (7) of Table 2B repeats the re-analysis of the Bertrand and Mullainathan (2004) data from Neumark (2012). In this case, the evidence of discrimination gets a bit stronger, and the variance of the unobservable is estimated to be larger for blacks. These findings are consistent with low quality resumes generating a bias against finding discrimination, although the qualitative conclusions are unchanged.

Thus, the conclusion from our re-examination of the labor market experiments is that the findings from the existing studies of discrimination against ethnic, racial, or sexual orientation minorities are not always robust to addressing the Heckman-Siegelman critique. All 14 estimates based on the existing studies, using the conventional approach, point to evidence of discrimination. But only six (or just under one-half) of the corrected estimates provide evidence of discrimination.[26]

This conclusion that the analysis of data from field experiments on labor market discrimination is not always robust is echoed in the findings reported in Neumark et al. (2016). They study age discrimination in hiring, and find that the evidence of discrimination against older women is robust to addressing the Heckman-Siegelman critique, but the evidence of discrimination against older men is not robust. On the other hand, some other recent papers using this technique do not find large differences. Carlsson et al. (2013) re-examine data from four previous studies of the Swedish labor market, each of which includes some form of the data required to implement the bias correction. Their re-analysis does not lead to large changes in the estimates of discrimination, although sometimes the estimated discrimination (against those with Arabic names, and in favor of women) becomes smaller. Three recent studies by Baert, all on the

[26] This includes the evidence from Carlsson and Rooth (2007), Drydakis (2014, for both gays and lesbians), Oreopoulos (2011, for Chinese and Indian), and Bertrand and Mullainathan (2004, significant at 10-percent level).

Belgian labor market, found no change in the estimates of discrimination in these experimental studies. Baert (2015) implemented this method in a study of discrimination against Turkish school-leavers in Belgium, using information on distance from the worker's residence to the workplace and other application characteristics to identify the heteroscedastic probit model, and report that this correction does not alter the conclusions. Baert (2014) applied the bias correction in an investigation of discrimination based on sexual orientation and family responsibilities, and found no bias or difference in reported results (Baert, 2014, footnote 15, p. 551). Baert (2016) found similar results in a study of hiring discrimination against disabled individuals (see pp. 83-84). Nunley et al. (2015) studied racial discrimination in hiring of recent college graduates in the United States. Applying the bias correction to their finding of a significant, lower interview rate to black graduates indicated that the baseline estimate of discrimination was understated, although the resulting estimated marginal effects through the level and variance were not statistically significant (p. 1118). Thus, among these latter studies, there is again sometimes an indication that the results are not robust to addressing the Heckman-Siegelman critique, although there is less clear of an indication that ignoring this critique leads to overstating discrimination.

Housing market field experiments

The results from the re-examination of the evidence from the housing discrimination studies are presented in Table 3. Ahmed et al. (2010) study discrimination against Arab applicants in Sweden, looking – as three of the four housing studies do – at both positive responses and offers of immediate showings. In this study, correcting for potential bias from differences in the variances of the unobservables does very little to change the conclusions. The estimates of lower positive responses or offers of immediate showings to Arab applicants become if anything more negative – most notably for immediate showing, where the estimate changes from −0.074 to −0.146 – and both estimates are statistically significant. The estimated effects of Arab ethnicity

through the variance are positive, and larger for immediate showings, corresponding to the larger negative estimate on the marginal effect through the level. The estimated variance of the unobservable is larger for Arab applicants, so combined, the estimates imply that the applications were lower quality than the population of applications to these landlords, biasing towards zero the conventional probit estimate of discrimination in immediate showings. Turning to the other diagnostics, in neither analysis are the overidentification restrictions, the restriction to a homoscedastic probit model, or equality of the standard deviations rejected. Thus, in this study evidence of discrimination persists.

These same conclusions are echoed in the remaining columns of the table – for the Bosch et al. (2010), Carlsson and Eriksson (2007), and Ewens et al. (2014) studies. In all cases, the bias-corrected estimates still lead to statistically significant evidence of discrimination based on race and ethnicity. And in most cases the point estimate for the marginal effect through the level is very close to the overall heteroscedastic probit estimate, while the estimates of the effect of race or ethnicity through the variance are very small.[27]

There is one case (Ewens et al., 2014) where the overidentifying restrictions are rejected at the 10-percent level (and the p-values for the other tests are fairly low). We therefore carried out an additional analysis, paralleling what we did with the Drydakis (2014) data on gay and straight male applicants. In this case, we estimated a less restrictive model that did not restrict the effects of percent black in the area or city to be the same across black and white applicants, based on the

[27] One reason for the robustness of the results in Carlsson and Eriksson (2013) could be because they use applications with substantial variation in applicant characteristics. The authors do this because by avoiding standardizing applicants to a very narrow range, the bias identified by the Heckman-Siegelman critique can be reduced, although this cannot ensure that the range of quality of actual applicants is not larger. It is also the case that – especially for the positive response outcome – the variances are nearly equal (the ratio of estimated standard deviations is 1.02), so that using a narrow range of applicant quality would not introduce bias.

estimates indicating that these interactions did not fit the pattern of equal coefficients in the latent variable model with probit coefficient differing because of differences in the variances of unobservables. In this case the overidentification restrictions were no longer rejected (the p-value was 0.877), yet the conclusions were similar to those in column (7) of Table 3. The overall estimate (standard error) of discrimination from the heteroscedastic probit model was −0.064 (0.023), and the unbiased estimated effect through the level was −0.067 (0.023).

Thus, the conclusion from our re-examination of the housing market studies is that the findings from the existing studies of discrimination against ethnic or racial minorities are robust to addressing the Heckman-Siegelman critique. With one minor exception, these past studies found evidence of discrimination, and our corrected estimates are qualitatively and usually quantitatively very similar.

Why might the housing market tests of callback for rental enquiries be more robust to addressing the Heckman-Siegelman critique? One possibility is that that the information provided in the housing market tests is sufficiently complete that there is little scope for a role for unobservables, and hence little impact of any differences in the variance of unobservables across groups. In housing markets, there may not be much more that matters to agents than ability to pay, and the information in the applications may convey this quite reliably. In contrast, an employer has an ongoing relationship with a worker, as do the employer's customers, so that many factors that are not conveyed in an on-line job application could potentially weigh on an employer's decision, and hence, correspondingly, differences in the variances of these unobserved factors across groups could matter much more.

6. Conclusions

The goal of this paper was to re-examine evidence from field experiments on labor market and housing market discrimination (experiments that, in general, identify the combined effect of

taste discrimination and statistical discrimination). Specifically, our goal is to see if the near-uniform findings of discrimination against minorities hold up after correcting for an important source of bias originally identified in Heckman and Siegelman (1993) – which we refer to as the "Heckman-Siegelman critique." This critique emphasises that even under quite ideal conditions for these studies, the evidence can be biased in either direction – or, equivalently, discrimination can be unidentified – if the variances of the unobservables differ across the groups studied. This is a plausible concern, given that a difference in the variances of unobservables across groups cannot be cannot be ruled out and indeed is at the core of early theoretical models of statistical discrimination (Aigner and Cain, 1977). We re-examine evidence from 10 studies that have the requisite data – applicant or other characteristics aside from the identifier for the group in question which shift the probability of call-backs or hires – implementing a correction for this bias proposed in Neumark (2012).

We find that for the housing market studies, the estimated effect of discrimination is robust to this correction. For the labor market studies, in contrast, the evidence is less robust; in about half of cases the estimated effect of discrimination either falls to near zero or becomes statistically insignificant, and in one case the sign changes.

We of course cannot definitively extrapolate from the 10 studies we were able to re-examine to the broader set of field experiments on discrimination by race, ethnicity, and sexual orientation. However, given that about half of the estimates of labor market discrimination that we could re-examine no longer provide statistical evidence of discrimination (or discrimination in the same direction) after correcting for bias from differences in the variance of unobservables, it seems reasonable to suggest that the overall (and overwhelming) evidence of labor market discrimination from field experiments is likely less robust than it seems. We have no doubt that in many countries there is discrimination in labor and housing markets against many groups, and that

– like the subset of studies we re-examine in this paper – the evidence of discrimination would frequently be robust to addressing the Heckman-Siegelman critique. But our evidence also indicates that in some cases a research design that enables a researcher to address this critique would not find evidence of labor market discrimination.

If nothing else, this conclusion implies that we need three types of research to draw more definitive conclusions from field experiments on labor and housing market discrimination: (1) more evidence using this kind of research design and methods; (2) more analysis of how best to implement these methods, what kinds of quality shifters provide the most informative estimates, etc.; and (3) further consideration of whether there are other ways to address the Heckman-Siegelman critique and whether they generate similar answers. Moreover, given the non-robustness of the experimental evidence on labor market discrimination, in particular, to addressing the Heckman-Siegelman critique, one could reasonably argue that future experimental studies of labor market discrimination (and perhaps of discrimination in any market) must take account of this critique to be regarded as credible.

References

Ahmed, A., Andersson, L. and Hammarstedt, M. (2010). 'Can discrimination in the housing market be reduced by increasing the information about the applicants?', *Land Economics*, 86(1), pp. 79-90.

Aigner, D. and Cain, G. (1977). 'Statistical theories of discrimination in labor markets', *Industrial and Labor Relations Review*, 30(2), pp. 175-87.

Altonji, J. and Blank, R. (1999). 'Race and gender in the labor market'. In *Handbook of Labor Economics, Volume 3*, edited by Orley C. Ashenfelter and David Card, 3143-259. Amsterdam: Elsevier.

Baert, S. (2014). 'Career Lesbians. Getting Hired for Not Having Kids?' *Industrial Relations Journal*, 45, pp. 543–561.

Baert, S. (2015). 'Field experimental evidence on gender discrimination in hiring: Biased as Heckman and Siegelman predicted?', *Economics*, 9, August 20, http://dx.doi.org/10.5018/economics-ejournal.ja.2015-25.

Baert, S. (2016). 'Wage Subsidies and Hiring Chances for the Disabled: Some Causal Evidence', *European Journal of Health Economics*, 17, pp. 71–86.

Baert, S. and Balcaen, P. (2013). 'The Impact of Military Work Experience on Later Hiring Chances in the Civilian Labour Market. Evidence from a Field Experiment'. *Economics: The Open-Access, Open-Assessment E-Journal*, 7, http://www.economics-ejournal.org/economics/journalarticles/2013-37.

Baert, S., Cockx, B., Gheyle, N. and Vandamme, C. (2015). 'Is there less discrimination in occupations where recruitment is difficult?', *Industrial and Labor Relations Review*, 68(3), pp. 467-500.

Baert, S. and Verhofstadt, E. (2015). 'Labour market discrimination against former juvenile delinquents: Evidence from a field experiment', *Applied Economics*, 47, pp. 1061-72.

Bertrand, M., Chugh, D., and Mullainathan, S. (2005). 'Implicit discrimination', *American Economic Review Papers and Proceedings*, 95(2), pp. 94-8.

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

Bosch, M., Carnero, M. and Farré, L. (2010). 'Information and discrimination in the rental housing market: Evidence from a field experiment', *Regional Science and Urban Economics*, 40(1), pp. 11-19.

Carlsson, M. and Eriksson, S. (2014). 'Discrimination in the rental market for apartments', *Journal of Housing Economics*, 23, pp. 41-54.

Carlsson, M., Fumarco, L. and Rooth, D.-O. (2013). 'Artifactual evidence of discrimination in correspondence studies? A replication of the Neumark method', IZA Discussion Paper No. 7619. Bonn, Germany: IZA.

Carlsson, M. and Rooth, D.-O. (2007). 'Evidence of ethnic discrimination in the Swedish labor market using experimental data', *Labor Economics*, 14(4), pp. 716-29.

Cornelißen, T. (2005). 'Standard errors of marginal effects in the heteroskedastic probit model', Institute of Quantitative Economic Research, Discussion Paper No. 230. Hanover, Germany: University of Hanover.

Dickinson, D. and Oaxaca, R. (2009). 'Statistical discrimination in labor markets: An

experimental analysis', *Southern Economic Journal*, 71(1), pp. 16-31.

Doleac, J. Stein. L.C.D. (2013). 'The Visible Hand: Race and Online Market Outcomes', *Economic Journal*, 123(572), pp. F469-92.

Drydakis, N. (2014). 'Sexual orientation discrimination in the Cypriot labor market: Distastes or uncertainty?' *International Journal of Manpower*, 35(5), pp. 720-44.

Ewens, M., Tomlin, B. and Wang, L.-C. (2014). 'Statistical discrimination or prejudice? A large sample field experiment', *Review of Economics and Statistics*, 96(1), pp. 119-34.

Heckman, J. (1998). 'Detecting discrimination', *Journal of Economic Perspectives*, 12(2), pp. 101-16.

Heckman, J. and Siegelman, P. (1993). 'The Urban Institute audit studies: Their methods and findings'. In *Clear and Convincing Evidence: Measurement of Discrimination in America*, edited by Michael Fix and Raymond J. Struyk, 187-258. Washington, D.C.: The Urban Institute Press.

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

Lee, H.A. and Khalid, M.A. (2016). 'Discrimination of high degrees: race and graduate hiring in Malaysia', *Journal of the Asia Pacific Economy*, 21(1), pp. 53-76.

Mincer, J. (1974). *Schooling, Experience, and Earnings*. New York: Columbia University Press.

Neumark, D. (2016). 'Experimental research on labor market discrimination.' NBER Working Paper No. 21262. Cambridge, MA: NBER.

Neumark, D. (2012). 'Detecting discrimination in audit and correspondence studies', *Journal of Human Resources*, 47(4), pp. 1128-157.

Neumark, D., Burn, I. and Button, P. (2016). 'Experimental age discrimination evidence and the Heckman critique.' *American Economic Review*, 106, pp. 303–308.

Nunley, J., Pugh, A., Romero, N., Seals, R. (2015). 'Racial Discrimination in the Labor Market for Recent College Graduates: Evidence from a Field Experiment', *B.E. Journal of Economic Analysis and Policy*, 15, pp.1093–1125.

OECD. (2013). *International Migration Outlook 2013*. Paris: OECD.

Oreopoulos, P. (2011). 'Why do skilled immigrants struggle in the labor market? A field experiment with thirteen thousand resumes', *American Economic Journal: Economic Policy*, 3(4), pp. 148-71.

Pager, D. (2007). "The use of field experiments for studies of employment discrimination: Contributions, critiques, and directions for the future', *The Annals of the American Academy of Political and Social Science*, 609(1), pp. 104-33.

Quillian, L., Pager, D., Hexel, O. and Midtboen, A. (n.d.). 'The persistence of racial discrimination: A meta-analysis of field experiments in hiring since 1972', unpublished paper.

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

Rich, J. (2014). 'What do field experiments of discrimination in markets tell us? A meta analysis of studies conducted since 2000', IZA Discussion Paper No. 8584. Bonn, Germany: IZA.

Zschirnt, E. and Ruedin, D. (2015). 'Ethnic discrimination in hiring decisions: A meta-analysis of correspondence tests 1990-2015', unpublished paper.

Table 1: Experimental Studies of Discrimination in Labor and Housing Markets Re-examined

Study (1)	Country (2)	Years (3)	Minority (4)	Outcome (5)	Majority call-back rate (6)	Estimated differential for minority (7)
A. Labor market field experiments						
Baert et al. (2015)	Belgium	2011-12	Turkish	Call-back	.329	-.082
						(.034)
				Immediate interview	.190	-.056
						(.026)
Carlsson and Rooth (2007)	Sweden	2005-6	Middle Eastern	Call-back	.269	-.095
						(.009)
Drydakis (2014)	Cyprus	2010-11	Gay	Call-back	.554	-.410
						(.010)
			Lesbian	Call-back	.523	-.411
						(.011
Lee and Khalid (2016)	Malaysia	2011	Malay (vs. Chinese)	Call-back	.221	-.152
						(.018)
Oreopoulos (2009)	Canada	2008	Chinese	Call-back	.142	-.053
						(.007)
			Indian	Call-back	.142	-.056
						(.007)
			Chinese-Canadian	Call-back	.142	-.063
						(.008)
			Pakistani	Call-back	.142	-.073
						(.009)
			Greek	Call-back	.142	-.035
						(.017)
			British	Call-back	.142	-.031
						(.011)
Bertrand and Mullainathan (2004)	United States	2001-2	Black-sounding names	Call-back	.097	-.030
						(.006)
B. Housing market field experiments						
Ahmed et al. (2010)	Sweden	2008	Arab/Muslim	Positive response	.514	-.171
						(.033)
				Immediate showing	.254	-.091
						(.024)
Bosch et al. (2010)	Spain	2009	Moroccan immigrants	Positive response	.590	-.133
						(.014)
				Immediate showing	.541	-.135
						(.014)
Carlsson and Eriksson (2014)	Sweden	2010-11	Arab	Positive response	.387	-.130
						(.012)
				Immediate showing	.271	-.110
						(.011)
Ewens et al. (2014)	United States	2009	Black	Positive response	.503	-.090
						(.019)

Note: All studies are correspondence studies. Column (7) reports marginal effect from probit models, our estimates, from following tables. In the Oreopoulos study, "Chinese-Canadian" means there was an English first name.

Table 2A: Estimates for Labor Market Discrimination Studies: Full Specifications

Study	Baert et al. (2015), Belgium		Carlsson and Rooth (2007), Sweden	Drydakis (2014), Cyprus		Lee and Khalid (2016) Malaysia
Outcome	Call-back	Immed. interview	Call-back	Call-back		Call-back
Minority group	Turkish, males		Middle Eastern, males	Gay	Lesbian	Malay
	(1)	(2)	(3)	(4)	(5)	(6)
A. Estimates from basic probit						
Minority, marginal effect	-.082	-.056	-.095	-.410	-.411	-.152
	(.034)	(.026)	(.009)	(.010)	(.011)	(.018)
B. Heteroscedastic probit model						
Minority, marginal effect	**-.096**	**-.072**	**-.095**	**-.384**	**-.304**	**-.201**
	(.034)	**(.028)**	**(.009)**	**(.040)**	**(.091)**	**(.038)**
Marginal effect through level (unbiased)	**.044**	**.073**	**-.102**	**-.476**	**-.499**	**.244**
	(.068)	**(.087)**	**(.023)**	**(.029)**	**(.016)**	**(.108)**
Marginal effect through variance	-.141	-.145	.007	.093	.195	-.445
	(.065)	(.093)	(.026)	(.065)	(.104)	(.142)
Standard deviation of unobservables, minority/non-minority	**.49**	**.55**	**1.03**	**1.59**	**2.27**	**.11**
Wald test, overidentification, ratios of coefficients equal (p-value)	.97	.93	.87	.09	.64	.94
LR test: standard vs. heteroscedastic probit (p-value)	.01	.10	.80	.06	.01	.01
Wald test, ratio of standard deviations = 1 (p-value)	.00	.03	.79	.18	.16	.00
Controls (job or applicants)	High education, over-educated, distance, vacancy duration, vacancies/unemployed, unemployment, % foreign, % Turkish, city, multiple jobs, average occupation wage, job quality, intensive/moderate customer contact		Unemployment spells, cultural activities, sport, personality, summer experiences, U.S. high school, high education, multiple employers, occupation	Enhanced cover letters, enhanced reference letters, first applicant, resume type, reference type, tester, occupation		Occupation, cover letter with good English, extracurricular skills, BA from private university, grades, language and writing skills stated (Malay, Chinese), MS Office, software/accounting skills, high quality CV, degree or degree project on CV, pre-university institution, job ad stated race
Clustered (within-pair design)	Yes		Yes	Yes		
N	736	736	5,636	4,846	4,194	3,009

Note: In Panel A, the marginal effect is based on the standard formula for a discrete variable, with other variables set at sample means. In Panel B, the continuous approximation for marginal effects is used, with the decomposition in equation (8) immediately below. The standard errors for the two components of the marginal effects are computed using the delta method. The only individual controls for which interactions are not introduced are for other demographic groups.

Table 2B: Estimates for Labor Market Discrimination Studies: Full Specifications

Study	Oreopoulos (2011), Canada						Bertrand and Mullainathan (2004), U.S.
Outcome	Call-back						Call-back
Minority group	Chinese	Indian	Chinese-Canadian	Pakistani	Greek	British	Black-sounding names
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A. Estimates from basic probit							
Minority, marginal effect	-.053	-.056	-.063	-.073	-.035	-.031	-.030
	(.007)	(.007)	(.008)	(.009)	(.017)	(.011)	(.006)
B. Heteroscedastic probit model							
Minority, marginal effect	**-.046**	**-.050**	**-.068**	**-.083**	**-.066**	**-.038**	**-.026**
	(.009)	**(.008)**	**(.009)**	**(.014)**	**(.073)**	**(.013)**	**(.007)**
Marginal effect through level (unbiased)	**-.131**	**-.101**	**-.029**	**-.076**	**-.169**	**.031**	**-.070**
	(.046)	**(.041)**	**(.054)**	**(.078)**	**(.208)**	**(.045)**	**(.040)**
Marginal effect through variance	.086	.052	-.040	-.007	.102	-.068	.045
	(.052)	(.046)	(.054)	(.070)	(.139)	(.052)	(.043)
Standard deviation of unobservables, minority/non-minority	**1.46**	**1.26**	**.84**	**.97**	**1.54**	**.75**	**1.26**
Wald test, overidentification, ratios of coefficients equal (p-value)	.72	.85	.78	.48	.66	.20	.42
LR test: standard vs. heteroscedastic probit (p-value)	.07	.22	.46	.92	.33	.21	.26
Wald test, ratio of standard deviations = 1 (p-value)	.19	.32	.42	.92	.55	.13	.37
Controls (job or applicants)	Extracurricular activities, top-ranked Bachelor's, Master's, occupation, speaking/social/writing skills required, female						Bachelor's, experience and square, volunteer, military service, email address, gaps in work history, work during school, academic honors, computer and other skills, female; in zip code (% high school dropout, college graduate, black, and white, log median household income)
Clustered (within-pair design)	Yes						Yes
N	5,866	6,373	4,468	3,978	3,388	3,934	4,784

Note: In Panel A, the marginal effect is based on the standard formula for a discrete variable, with other variables set at sample means. In Panel B, the continuous approximation for marginal effects is used, with the decomposition in equation (8) immediately below. The standard errors for the two components of the marginal effects are computed using the delta method. The only individual controls for which interactions are not introduced are for other demographic groups. Some skills are specific to immigrant groups and used to distinguish among immigrants (such as specific language fluencies or where experience obtained), and are not included.

Table 3: Estimates for Housing Discrimination Studies

Study	Ahmed et al. (2010), Sweden		Bosch et al. (2010), Spain		Carlsson and Eriksson (2007), Sweden		Ewens et al. (2014), U.S.
Outcome	Positive response	Immediate showing	Positive response	Immediate showing	Positive response	Immediate showing	Positive response
Minority group	Arab/Muslim		Moroccan immigrants		Arabic/Muslim		Black
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A. Estimates from basic probit							
Minority, marginal effect	-.171	-.091	-.133	-.135	-.130	-.110	-.090
	(.033)	(.024)	(.014)	(.014)	(.012)	(.011)	(.019)
B. Heteroscedastic probit model							
Minority, marginal effect	**-.165**	**-.074**	**-.136**	**-.136**	**-.131**	**-.113**	**-.089**
	(.034)	**(.027)**	**(.017)**	**(.017)**	**(.013)**	**(.011)**	**(.019)**
Marginal effect through level (unbiased)	**-.182**	**-.146**	**-.136**	**-.135**	**-.134**	**-.074**	**-.092**
	(.035)	**(.049)**	**(.018)**	**(.015)**	**(.026)**	**(.034)**	**(.019)**
Marginal effect through variance	.017	.072	.001	-.001	.004	-.039	.003
	(.019)	(.058)	(.004)	(.014)	(.025)	(.035)	(.003)
Standard deviation of unobservables, minority/non-minority	**1.20**	**1.35**	**.91**	**.98**	**1.02**	**.85**	**1.08**
Wald test, overidentification, ratios of coefficients equal (p-value)	.59	.91	.33	.52	.87	.93	.07
LR test: standard vs. heteroscedastic probit (p-value)	.32	.20	.74	.95	.88	.26	.18
Wald test, ratio of standard deviations = 1 (p-value)	.37	.29	.74	.95	.89	.22	.20
Controls (area or applicants)	Enhanced application, rent, space, rooms, metro, company		Enhanced application, rent, rooms, urban, company, female		Jobs, exercise, nightclub, smoker, references, female, age		Mother's estimated education, positive email, negative email, rent, relative rent, rent in area, one BR, cost, % male, % black in area/city, female, Muslim name
Clustered (within-pair design)	No		Yes		No		No
N	959	959	4,716	4,716	5,827	5,827	13,800

Note: In Panel A, the marginal effect is based on the standard formula for a discrete variable, with other variables set at sample means. In Panel B, the continuous approximation for marginal effects is used, with the decomposition in equation (8) immediately below. The standard errors for the two components of the marginal effects are computed using the delta method. The only individual controls for which interactions are not introduced are for other demographic groups.

E-lit-007: Systemic Discrimination Among Large U.S. Employers

Patrick M. Kline, Evan K. Rose and Christopher R. Walters | Quarterly Journal of Economics 137(4): 1963-2036 (NBER Working Paper 29053) | 2022

https://www.nber.org/system/files/working_papers/w29053/w29053.pdf

About this source

Applications carrying distinctively Black names were contacted 2.1 percentage points less often than otherwise identical applications with distinctively white names (25.1 percent versus 23.0 percent), a difference significant at p < 10^-32. That is a 9 percent proportional penalty - far below the roughly 36 percent white advantage Quillian et al. (2017) report for recent U.S. correspondence experiments, and far below Bertrand and Mullainathan's 50 percent. The authors are explicit that this is arithmetic, not a change in employer behavior in levels: the level gap of about 2 percentage points sits squarely alongside Nunley et al. (2015) at 2.6 points and Agan and Starr (2018) at 2.4 points, and the proportional gap is smaller only because these very large employers contacted 24 percent of all applicants, nearly three times Bertrand and Mullainathan's rate. They offer one substantive reason it could also be genuinely smaller - very large firms may discriminate less, as a Canadian experiment found. The penalty is stable rather than fading: white applicants were contacted more often in 12 of 13 study months, and constancy across the five waves cannot be rejected. The average gap conceals enormous firm heterogeneity - a between-firm standard deviation of 1.9 percentage points, almost as large as the mean - with the top quintile of firms responsible for 46 percent of all lost contacts, a Gini of 0.40, and 23 individual companies identifiable as discriminators while controlling the false discovery rate at 5 percent. By contrast the average gender gap is zero, but the between-firm standard deviation of 2.7 points shows some firms favor men and others women, and applicants over 40 face a 0.6 point penalty.

Bound as the original PDF. Retrieved 2026-09-10. sha256 015e29ffa702a9f0...

Bound into an Economic Briefing · by Claude, may contain mistakes

Full document begins on the next page.

NBER WORKING PAPER SERIES

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, Revised May 2022

We thank Sendhil Mullainathan, Sapna Raj, and Jenny Yang for helpful conversations, our discussants Peter Hull and Peter Q. Blair, and seminar participants at UC Berkeley, Chicago Booth, the University of Chicago HCEO Working Group, the University of Wisconsin, CEPR, IZA, the NBER Labor Studies meetings, 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, Elena Marchetti-Bowick, 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.

© 2021 by Patrick M. Kline, Evan K. Rose, and Christopher R. Walters. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Systemic Discrimination Among Large U.S. Employers
Patrick M. Kline, Evan K. Rose, and Christopher R. Walters
NBER Working Paper No. 29053
July 2021, Revised May 2022
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 discrimination against distinctively Black names 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

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

Evan K. Rose
Becker Friedman Institute
University of Chicago
Chicago, IL 60637
and NBER
ekrose@uchicago.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. Yet a large social science literature analyzing resume correspondence experiments finds 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). In a reanalysis of several correspondence experiments, Kline and Walters (2021) find that discriminatory biases vary tremendously across job vacancies. Less is known, however, about the extent to which discriminatory jobs are concentrated in particular companies. 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. First, 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 employers. Second, if the most biased firms also tend to offer the highest wages, the contribution of discrimination to observed disparities will tend to be amplified (Card, Cardoso and Kline, 2016; Gerard et al., 2021). Third, if only a few firms discriminate, and do so heavily, it may be possible for government regulators to target these companies for audits and investigations. For instance, the Office of Federal Contract Compliance (OFCCP) annually audits thousands of federal contractors for compliance with equal employment laws (Maxwell et al., 2013). Likewise, the U.S. Equal Employment Opportunity Commission (EEOC) routinely launches investigations into whether particular companies have engaged in "systemic discrimination," a term they define 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*).

This paper reports the results of a massive nationwide correspondence experiment designed to measure patterns of discrimination by large U.S. companies. The twin goals of our analysis are to quantify the extent to which discriminatory patterns differ across firms and to assess the feasibility of using experimental evidence to target firms likely to be engaged in discrimination. To facilitate these goals, our experiment was designed to repeatedly elicit signals of bias from specific companies. Unlike traditional audit studies that passively sample jobs from newspapers or job boards (e.g., Bertrand and Mullainathan, 2004), we prospectively applied to entry level job vacancies hosted on the web portals of 108 Fortune 500 firms. For each company, we sampled up to 125 entry-level

jobs in distinct U.S. counties. 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 consistent organization-wide patterns.

Following a large social science literature (Bertrand and Duflo, 2017; Baert, 2018), our 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. Over 83,000 job applications were sent in total, providing uniquely precise signals of employer conduct.

Overall, 24% of the applications we sent were contacted by employers within 30 days. This contact rate is nearly three times greater than what Bertrand and Mullainathan (2004) found in their seminal experiment, suggesting that our fictitious applicants were viewed as plausible job candidates by employers. 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. Past work has typically found larger proportional effects, which may be attributable both to less biased behavior among the extremely large employers we study and the high overall contact rates yielded by our experiment.

A key finding of our analysis is that patterns of discrimination against Black names vary substantially across employers. 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. Despite this wide variability, we cannot reject the null hypothesis that all 108 firms in our experiment weakly favor white names. An application of Efron's (2016) empirical Bayes (EB) deconvolution estimator reveals that while most firms exhibit mild discrimination against Black applicants a few exhibit very large biases. We estimate that the top quintile of discriminating firms are responsible for nearly half of the lost contacts to Black applicants in our experiment. The Gini coefficient of employer contact gaps is estimated to be approximately 0.4, suggesting that discrimination against Black names is roughly as concentrated among firms in our experiment as income is among U.S. households.

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 large U.S. employers exhibit corresponding cross-company patterns of heterogeneity in their average gender contact gaps. Like racial discrimination, gender discrimination is highly concentrated in particular firms, with the top quintile of discriminating firms responsible for nearly 60% of contacts lost to gender discrimination and a Gini concentration coefficient of roughly 0.5.

While our main focus is on race and gender, we also assess the extent of discrimination on several other dimensions. A modest contact penalty of 0.6 percentage points is found 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.

Surprisingly, geographic variation in race, gender, and age discrimination is relatively muted. We cannot reject the null hypothesis that mean contact gaps for gender and age are equal across all 50 states, and find only marginally significant evidence against this null for racial contact gaps. In contrast, 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, suggesting that screening algorithms are unlikely to drive the firm differences we measure.

Consistent with classic models of customer discrimination, both racial and gender contact gaps are estimated to be larger in sectors intensive in jobs requiring social interaction. In line with the predictions of Becker (1957), racial contact gaps are smaller at more profitable firms. Racial contact gaps also tend to be smaller among federal contractors, which is consistent with Miller (2017)'s finding that contracting with the federal government yields sustained increases in Black employment. 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 organization wide biases.

The finding of significant employer heterogeneity in discriminatory conduct motivates an investigation of which particular organizations are likely violating the Civil Rights Act. As a first approach to characterizing detection possibilities, we form EB posterior mean estimates of the contact gap at each individual firm. Firms with posterior mean contact gaps in the top quartile of the distribution are estimated to account for roughly half of the contacts lost to racial discrimination. Discrimination is disproportionately clustered in customer facing sectors including the auto services and sales sector and certain forms

of retail. We find large posterior mean contact gaps favoring women at apparel stores and slightly less pronounced gaps 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 result implies that at least 22 of these 23 firms should be expected to exhibit 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 detect any firms that discriminate against women when 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 conclude are likely 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, indicating a nationwide pattern of discrimination against Black names.

We conclude with an economic analysis of optimal auditing strategies meant to mimic the objectives and constraints of regulatory authorities such as the EEOC or OFCCP. Building on the framework introduced in Kline and Walters (2021), 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 23 firms we classify as discriminating against Black names are estimated to account for nearly 40% of lost contacts to Black applicants in our experiment.

Our findings demonstrate that it is possible to target the individual firms responsible

for a substantial share of discrimination against Black names while maintaining a tight limit on the expected number of false positives. The evidence of discriminatory patterns uncovered here can, in principle, be used by organizations such as the EEOC or OFCCP to target audits and investigations more effectively. Alternatively, this information can be shared directly with the firms themselves, or even made public, potentially enabling companies to preemptively reform their practices, perhaps by adopting the recruiting policies of their less discriminatory peers.

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-qualified 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 disparate treatment, however motivated. The correspondence experiment we study was designed to manipulate employer percep-

[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`).

tions 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] Although it is unclear whether the statistical evidence provided in an audit study would, on its own, be sufficient to successfully litigate a Title VII disparate treatment claim, such evidence may be helpful in building a case or in targeting investigations that lead to the discovery of additional evidence that eventually proves decisive.[3] Conversely, correspondence evidence suggesting equal treatment of workers with different characteristics could, in principle, be used by firms to *counter* charges of employment discrimination. However, further evidence would likely be required for such a determination, as audit studies may fail to detect biases that manifest only at later stages of the hiring process or among applicants with qualification levels outside those considered in the study.

While the social science literature has proposed several distinct theories and definitions of systemic discrimination (e.g., Pincus, 1996; Reskin, 2012), our usage of this phrase is motivated by the EEOC's definition of this term as a "pattern or practice" of discrimination (U.S. EEOC, 2006*b*; Kim, 2015). 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.[4] 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 discrimination against protected characteristics 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.[5] Evidence

[2]For discussion of the potential legal ramifications of handling fictitious applications based upon racial perceptions of names, 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). 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]Explicit evidence of intent to discriminate is *not* required to establish a prima facie case for disparate treatment. The EEOC's own guidance states that "discriminatory motive can be inferred from the fact that there were differences in treatment" (International Brotherhood of Teamsters v. United States, 431 U.S. 324 (1977)). In some cases, large statistical disparities alone can also constitute prima facie evidence of intentional discrimination (Hazelwood School Dist. v. United States, 433 U.S. 299 (1977)).

[4]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).

[5]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

of patterns of discrimination by federal contractors is especially pertinent to the OFCCP, which has broad discretion to audit contractors for compliance with executive orders prohibiting employment discrimination and regularly levies fines and, in some cases, even debars contractors when violations are found (Maxwell et al., 2013).

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).[6] 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. A finding that effects of this nature are present across a large set of establishments suggests 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 sufficient geographic variation and routinely posted enough entry-level jobs on an easily-accessible portal to satisfy our sampling criteria. These 108 large firms, 10 of which are subsidiaries of parent companies in the Fortune 500, employed roughly 15 million workers in 2020 according to Compustat and cover a wide array of industries detailed later on in Table 10.

We sampled 125 entry-level job vacancies from each employer, with each vacancy corresponding to an establishment in a different U.S. county. Sampling was organized in

landlords (Fang, Guess and Humphreys, 2019) and employers (Pager, Bonikowski and Western, 2009).

[6]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."

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 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.[7] 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 that of 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% 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.[8] 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 uni-

[7]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.

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

formly 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 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.[9]

[9]The pre-analysis plan is stored in the AEA RCT registry with number AEARCTR-0004739.

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 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 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 reports 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 cat-

egory ($p \geq 0.24$). 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. Of course, differences in employer treatment of distinctively Black and white (or male and female) names may, in part, reflect stereotypes about average productivity differences between these groups. We leave an investigation of this possibility to future work. Note however that the courts—not to mention potential customers, employees, and corporate shareholders—are likely to view claims that an employer discriminates against applicants with Black (or female) names based on productivity grounds as a pretext for illegal discrimination.

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 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 esti-

mates 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 the effects of LGBTQ club membership 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 at the 10% level ($p = 0.065$), 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

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

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, 1988; Aaronson, Barrow and Sander, 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 Testing for firm components

To test formally for the significance of firm-level contact gap variation, we report a 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. To improve power, we instead employ the high-dimensional moment inequality testing procedure of Bai, Santos and Shaikh (2021), which drops firms with strongly positive z-scores.

The first two columns of Table 4 report 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 suggests 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 result strongly suggests the presence of 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 ($p \leq 0.05$). These findings extend and corroborate recent work by Kline and Walters (2021) and Hangartner, Kopp and Siegenthaler (2021) who conclude that gender discrimination varies bi-directionally across jobs in Mexico and Switzerland, respectively.

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. By contrast, the null hypothesis that no firms discriminate against older applicants is rejected at conventional levels ($p = 0.03$).

6.3 Variance component 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. Graphical evidence of the cross-wave stability of race, gender, and age contact gaps is provided in Appendix Figure A6, which plots firm contact gaps in each wave against their leave-wave-out means. These plots also reveal that firm contact gaps for race and gender are not significantly correlated with each other.

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.

6.4 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 0.8 percentage points, less than half the magnitude of the between-firm standard deviation reported in Table 4.

Controlling for firm fixed effects reduces the modest state variation in contact gaps even further. Table 7 uses the leave-out estimator 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 $(0.190/0.267)^2 \times 100 = 51\%$ 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.[10]

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 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.[11] 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

[10]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.

[11]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.

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. In unreported results, we also found no significant difference in contact gaps between firms that required a battery of cognitive and personality tests and those that did not. The platforms themselves therefore do not appear to be an important driver of the between-firm differences we document.

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 nonetheless offer a low-dimensional summary of the sorts of jobs, establishments, and firms where discrimination tends to be more or less severe. Figures 3, 4, and 5 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 established that contact gaps vary substantially across job titles but that this variation is insignificant conditional on firm effects. While this finding suggests that variation in discrimination across job titles is mostly attributable to the identity of the parent firm, studying lower dimensional summaries of job titles may allow detection of more subtle relationships. A large literature (e.g., Deming, 2017; Hurst, Rubinstein and Shimizu, 2021) finds that the task content of work provides a useful summary of changes in the occupational structure of wages and employment. Adopting this approach, Figure 3 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 correlation may reflect employer concerns regarding customer discrimination, the quantitative importance of which has proven difficult to establish decisively (Holzer and Ihlanfeldt, 1998; Leonard, Levine and Giuliano, 2010; Hurst, Rubinstein and Shimizu, 2021). 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 ($p = 0.20$). This finding casts doubt on simplistic versions of the customer discrimination hypothesis where all employers discriminate differentially in customer facing jobs. For gender, the task content variables are marginally significant conditional on firm fixed effects ($p = 0.01$), suggesting that, at a typical large firm, men face discrimination in customer facing jobs while women face discrimination at jobs intensive in manual skills.

Appendix Figure A7 decomposes the relationship between contact gaps and job task content into within and between industry components. Within industry relationships between racial contact gaps and task content are weak and statistically insignificant, indicating that the task content correlations documented in Figure 3 are driven primarily by between industry variation. Contact gaps are especially strongly related to industry average customer interaction scores ($p = 0.001$). In contrast, the relationship between gender contact gaps and task content is strong both within and between industries. These results show that discrimination against both Black and male names is more intense in customer facing sectors, regardless of whether the job itself is customer facing. This finding may indicate that firms in different sectors tend to adopt different corporate cultures and human resources practices affecting all their jobs.

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 4 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.[12] It is worth noting, however, that many jobs in our sample did not specify an exact establishment address; consequently, 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 appears to be heightened 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 4, 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. Despite achieving statistical significance, these geographic correlations are all fairly modest in magnitude, which aligns with our earlier finding in Table 6 of a small but statistically significant between state variance component to racial discrimination.

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 4 shows that including firm fixed effects renders the establishment characteristics jointly insignificant ($p = 0.34$). Similar to our analysis of job titles, this finding 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. Consistent with our earlier finding in Table 6 of a negligible state component to gender gaps, Panel (b) of Figure 4 shows insignificant relationships between gender contact gaps and local demographics, measures of prejudice, and establishment charac-

[12]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.

teristics. 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 ($p \geq 0.35$).

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 5 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 5, 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.[13] 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 5, 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. As documented in Appendix Table C2 centralization is elevated among federal contractors ($p = 0.038$) but we cannot reject that it is unrelated to our other firm level predictors in a multivariate regression. Since this predictor is calculated using the outcome data 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,

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

1995) intended to avoid any mechanical relationship between job-level callback propensities and gaps. The negative coefficient estimate suggests that firms at which hiring responsibility is more centralized are less prone to bias, perhaps because rules replace the discretionary judgements of individual workers at firms with more sophisticated human resources practices. Overall, the firm-level variables in Figure 5 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 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 evaluating the sample average

$$\hat{g}_\Delta(x) = \frac{1}{108}\sum_{f=1}^{108}\frac{1}{s_f}\hat{g}_\mu(x/s_f).$$

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 and statistically insignificant relationships between standard errors and z-scores for both race and gender, suggesting that independence is a reasonable approximation. Appendix E explores three alternate approaches to modeling the joint distribution of s_f and z_f and shows that they yield results similar to those found when independence is imposed.

Panel (a) of Figure 6 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.[14] 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.[15] For comparison with the estimated densities, the background of Figure 6 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. Pointwise confidence intervals on the estimated densities are reported in Appendix Figure A10.

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

[14] 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.

[15] 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 A12.

discrimination is driven by a small group of highly discriminatory firms. Figure 7 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 contact gaps are even more concentrated, with firms in the top quintile responsible for 56% 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 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.[16]

10 Firm-Specific Estimates

The finding that discrimination is highly concentrated raises the question of whether it is possible to deduce the contact gaps of particular firms. Firm-specific estimates could, in principle, be shared with company executives, providing them with an assessment of their organization's biases, or with regulators to help them target audits or other enforcement efforts more effectively. Although the sample contact gaps $\hat{\Delta}_f$ provide unbiased estimates of the contact gap at each firm, those estimates are often quite noisy. 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 each specific firm.

10.1 Posterior mean 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

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

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 A11 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.

Appendix Figure A13 assesses the out of sample predictive power of these posterior means by shrinking z-scores constructing using only the first three waves of the experiment and comparing these shrunk values to contact gaps in the final two waves of the experiment. For race, we find a correlation between our EB predictions and the latent

contact gaps in the last two waves of 0.7, indicating very significant out of sample forecasting ability even when working with predictions that discard 40% of our microdata.

The posterior mean racial contact gaps vary systematically across industries. Figure 8 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 6, 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 9 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. These gender bias correlations align closely with the matched pair audit evidence reported by Neumark, Bank and Van Nort (1996) who find that women are discriminated against at upscale restaurants, which tend to pay high wages and to be male dominated, but are weakly preferred at lower price restaurants that tend to pay lower wages and to be female dominated.

One potential explanation for the divergent correlation patterns uncovered for sex and race in Figure 9 is that job seekers know that certain sectors (e.g. women's apparel) discriminate on the basis of gender, perhaps due to a mix of coworker and customer discrimination. This common knowledge allows workers to sort away from biased jobs, mitigating to some extent the burden of discrimination as in Becker's (1957) classic

model. Industry patterns of racial discrimination, by contrast, may be more difficult to discern, particularly if these patterns are driven by variation in opaque corporate recruiting protocols. When discriminatory patterns are not common knowledge, less pronounced sorting patterns will arise and a larger burden may fall 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 with degrees of freedom equal to the number of jobs at the firm minus one. Histograms of the resulting p-values for the null that firm specific contact gaps equal zero appear in Figure 10. 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 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. $FDR(p)$ 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

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

where $\widehat{FDR}(p)$ is an estimator of false discovery rates based on the empirical distribution of p-values.[18] If $\widehat{FDR}(p)$ were a consistent estimator of $FDR(p)$ 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 $FDR(p)$ 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

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

where $\lambda \in [0, 1)$ is a tuning parameter governing how deep to look in the right tail of empirical p-values. For any choice of λ, however, the probability limit of $\hat{\pi}_0(\lambda)$ will lie weakly above the true 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. To assess the degree of uncertainty in our estimate, we also report the upper limit of a non-parametric confidence interval for π_0 developed by Armstrong (2015).

[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 $\min_{t\geq\hat{p}_f} \widehat{FDR}(t)$, which ensures that q-values are non-decreasing for nested rejection thresholds.

10.3 Which firms discriminate?

Figure 10 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 10 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 $\hat{q}_f$.[19]

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.[20] 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 40% of firms may be discriminating against Black applicants, while under two-tailed tests the share discriminating may be as low as 30%.

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). The largest q-value in this set of firms is 0.047, so we should expect at most $23 \times 0.047 = 1.08$ false discoveries if these 23 firms are 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.

Though we expect at most 1 of the 23 firms with q-values below 0.05 to have racial contact gaps equal to zero, the actual number of false discoveries may differ from its expected value. To get a sense of how many false discoveries could potentially arise in an unfavorable scenario, Appendix Figure A14 plots the posterior distribution of false discoveries implied by the LFDRs of these 23 firms.[21] Reassuringly, the posterior probability mass function of false discoveries is tightly concentrated around its mean, with the

[19] 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.

[20] 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.

[21] The number of false discoveries exhibits a Poisson binomial posterior distribution with probabilities given by the LFDRs of the hypotheses under consideration. See Basu et al. (2021) for discussion.

posterior chances of three or more of these firms exhibiting contact gaps of zero being less than 2%.

The lower panels of Table 8 reveal that conclusions regarding the set of firms likely to be discriminating against Black names are remarkably robust to the method used to bound π_0. In fact, if we use randomization inference based p-values to estimate π_0, 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 8, 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 0.13, implying that at least 87% 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 A15 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 10 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 not close to one, a condition which seems to be violated 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

[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 $\hat{\pi}_0 = 1$. 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 $\phi_f < 1$.

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.

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 0.073. The mean gap across jobs is 0.020 with associated standard error of 0.002. Consequently, the lower bound prevalence is estimated to be $\frac{(0.020)^2-(0.002)^2}{(0.020)^2-(0.002)^2+(0.073)^2} \approx 0.07$, 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 12 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 $\hat{q}_f < 0.05$ 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 employer 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]

[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 as 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 $p \to \infty$, the auditor grows increasingly risk averse and U approaches U^e.

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 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 $\bar{\Delta}_f(G) = \mathbb{E}_G[\Delta_f|\mathcal{E}]$ 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 11 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. Since $\hat{G}_\Delta$ exhibits a mean contact gap of 2.1 percentage points (see Figure 6), 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 approximately 37% of lost contacts. Investigating the same share of firms based on posterior mean rankings would be expected to yield only an additional 4% of 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 the contacts lost to racial discrimination while maintaining a tight limit on the expected number of 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 concentration of discrimination among particular employers may amplify group disparities if discriminatory firms tend to offer higher wages. While we found no relationship at the industry level between racial wage gaps and racial contact gaps, industry contact gaps favoring men were found to be predictive of larger gender wage gaps. An interesting topic for future research is to assess the extent to which the firm-level contact gaps uncovered in this experiment correlate with group disparities in firm wage fixed effects such as those studied by Card, Cardoso and Kline (2016) or Gerard et al. (2021).

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. 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 EEOC maintains an internal target for the share of its litigation docket comprised of systemic discrimination cases. The appropriate level of this target has been a topic of recurring debate in Congress, where oversight committees have questioned the EEOC's

choice to prioritize systemic investigations of firms over individual level claims of discrimination (Kim, 2015). Our finding that discrimination is highly concentrated in particular companies lends some credence to the notion that appropriately targeted systemic investigations have the potential to remedy, and perhaps also prevent, discrimination affecting a wide swath of the labor force. Whether the targeting of systemic investigations by the EEOC, or of contractor audits by the OFCCP, can be improved with the adoption of statistical methods of the sort developed here is an open and interesting question.

Enforcement actions are inevitably costly and contentious. It is natural to wonder whether bias at the most discriminatory firms can be preemptively reduced or eliminated by modifying 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 (negative) predictors of both racial and gender contact gaps found in our correspondence experiment is callback centralization, which is notably elevated among federal contractors. This finding leads us to suspect that human resources practices play an important role in translating the biased judgements of individuals into biased behavior by organizations. While centralizing interview decisions might serve to reduce discrimination, such changes may also simply postpone discrimination to a later stage of the hiring 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).

[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."

References

Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. "Teachers and student achievement in the Chicago public high schools." *Journal of labor Economics*, 25(1): 95–135.

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.

Basu, Pallavi, Luella Fu, Alessio Saretto, and Wenguang Sun. 2021. "Empirical bayes control of the false discovery exceedance." *arXiv preprint arXiv:2111.03885*.

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.

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

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." *University of Chicago Legal Forum*, 139.

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.

Gerard, François, Lorenzo Lagos, Edson Severnini, and David Card. 2021. "Assortative matching or exclusionary hiring? the impact of employment and pay policies on racial wage differences in brazil." *American Economic Review*, 111(10): 3418–57.

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*.

Hangartner, Dominik, Daniel Kopp, and Michael Siegenthaler. 2021. "Monitoring hiring discrimination through online recruitment platforms." *Nature*, 589(7843): 572–576.

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.

Hurst, Erik, Yona Rubinstein, and Kazuatsu Shimizu. 2021. "Task-based discrimination." 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.

Kim, Pauline T. 2015. "Addressing systemic discrimination: Public enforcement and the role of the EEOC." *BUL Rev.*, 95: 1133.

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.

Koenker, Roger, and Ivan Mizera. 2014. "Convex optimization, shape constraints, compound decisions, and empirical Bayes rules." *Journal of the American Statistical Association*, 109(506): 674–685.

Koenker, Roger, and Jiaying Gu. 2017. "REBayes: an R package for empirical Bayes mixture methods." *Journal of Statistical Software*, 82: 1–26.

Krueger, Alan B, and Lawrence H Summers. 1988. "Efficiency wages and the inter-industry wage structure." *Econometrica: Journal of the Econometric Society*, 259–293.

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

Maxwell, Nan, Aravind Moorthy, Caroline Massad Francis, Dylan Ellis, et al. 2013. "Using Administrative Data to Address Federal Contractor Violations of Equal Employment Opportunity Laws." Mathematica Policy Research.

Miller, Conrad. 2017. "The persistent effect of temporary affirmative action." *American Economic Journal: Applied Economics*, 9(3): 152–90.

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.

Neumark, David, Roy J Bank, and Kyle D Van Nort. 1996. "Sex discrimination in restaurant hiring: An audit study." *The Quarterly journal of economics*, 111(3): 915–941.

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_2016OFC00006_(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.

Pincus, Fred L. 1996. "Discrimination comes in many forms: Individual, institutional, and structural." *American Behavioral Scientist*, 40(2): 186–194.

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.

Reskin, Barbara. 2012. "The race discrimination system." *Annual Review of Sociology*, 38: 17–35.

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.

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."

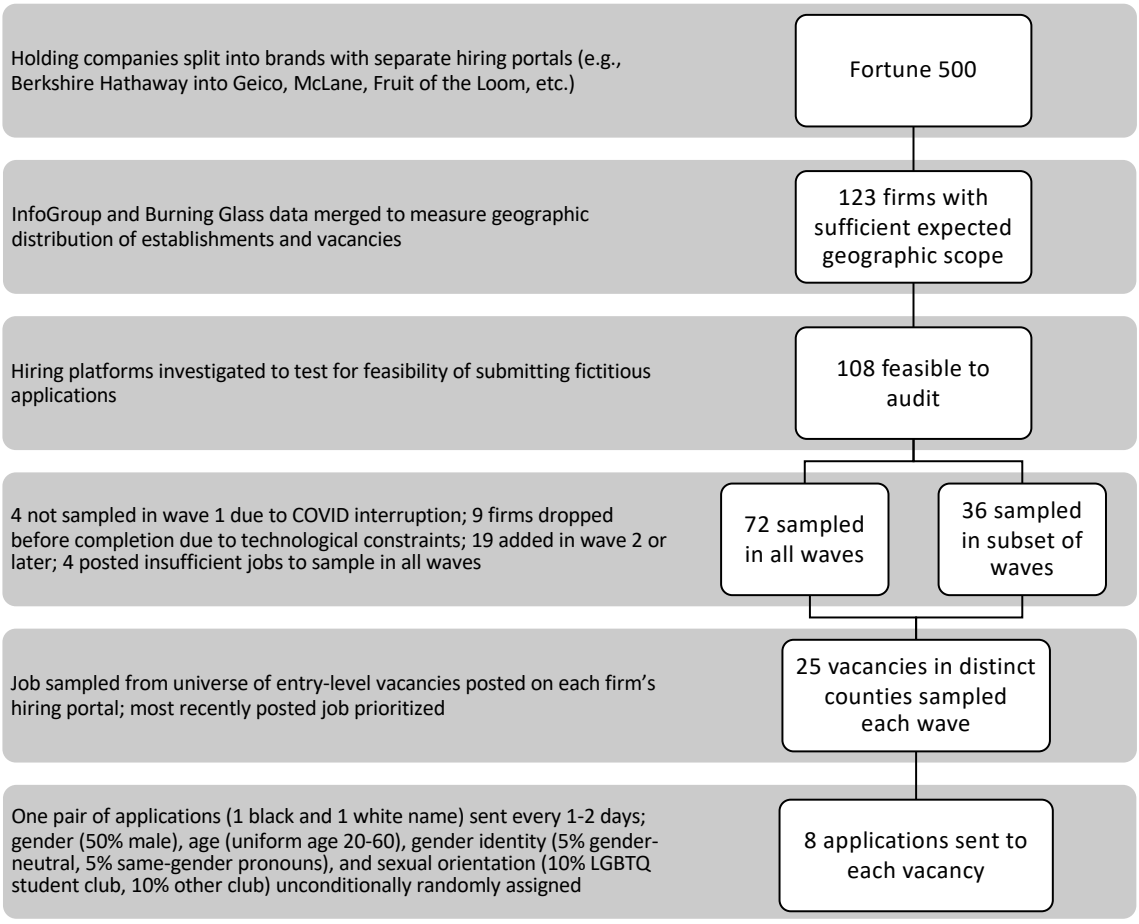
U.S. EEOC. 2019. "EEOC Sues Schuster for Sex Discrimination." *Press Release*, https://www.eeoc.gov/newsroom/eeoc-sues-schuster-sex-discrimination.

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.

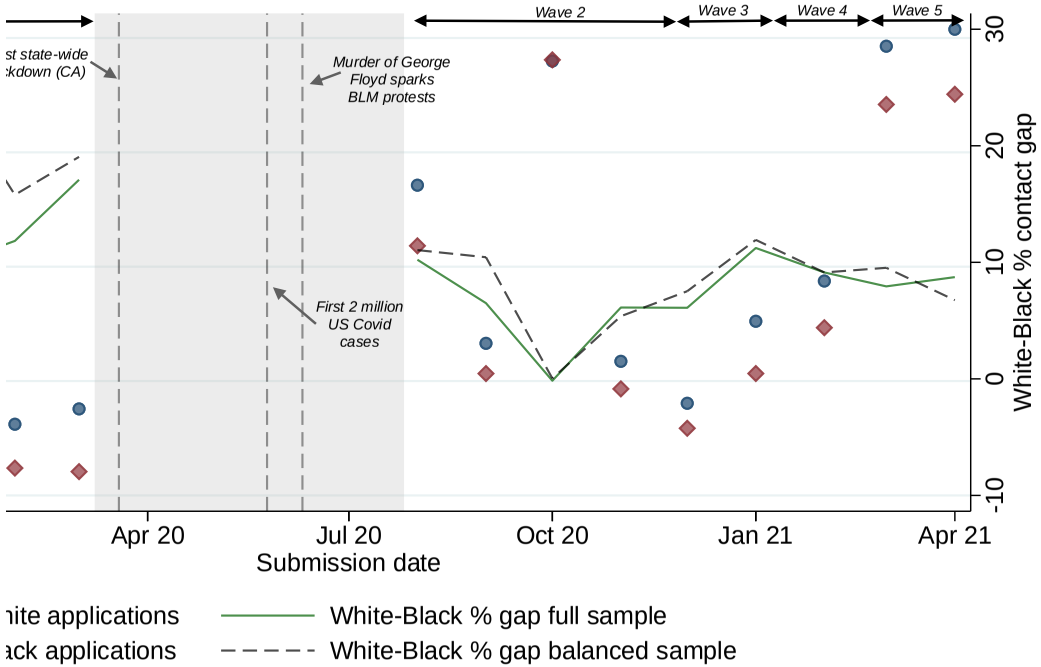
Figures

Figure 1: Overview of sampling strategy and experimental design



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



t rates for black and white applications by month and year of submission. The gray region corresponds to the ased due to Covid-19 related shut-downs. The green and dotted black lines plot the white-Black contact rate ack contact rate for the full and balanced samples. An F-test fails to reject that the white/Black percentage e same in all waves of the experiment ($F = 0.85$, $p = 0.50$).

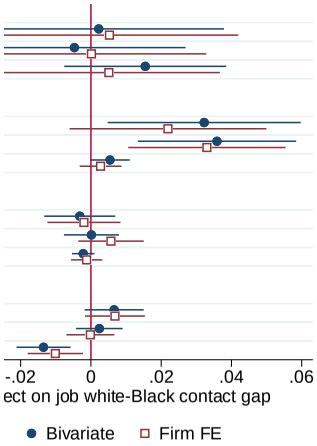
Figure 3: Relationships between contact gaps and job task content



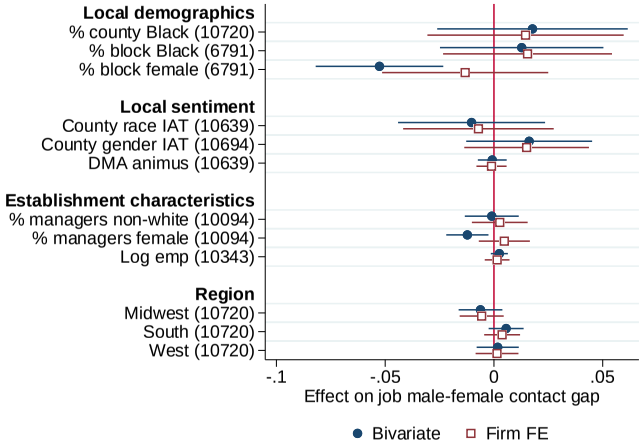
hip between O*Net measures of job-level task content and contact gaps for race and gender. Each relationship with job-level contact gaps as the outcome. All jobs with defined contact gaps for each attribute are included. on is in parentheses. Task measures are normalized to have standard deviation one in sample. "Bivariate" ons of contact gaps on the covariate alone. "Multivariate" points plot effects when all covariates are included confidence intervals based on robust standard errors. Appendix C provides a complete description of task

e 4: Relationships between contact gaps and establishment characteristics

ce

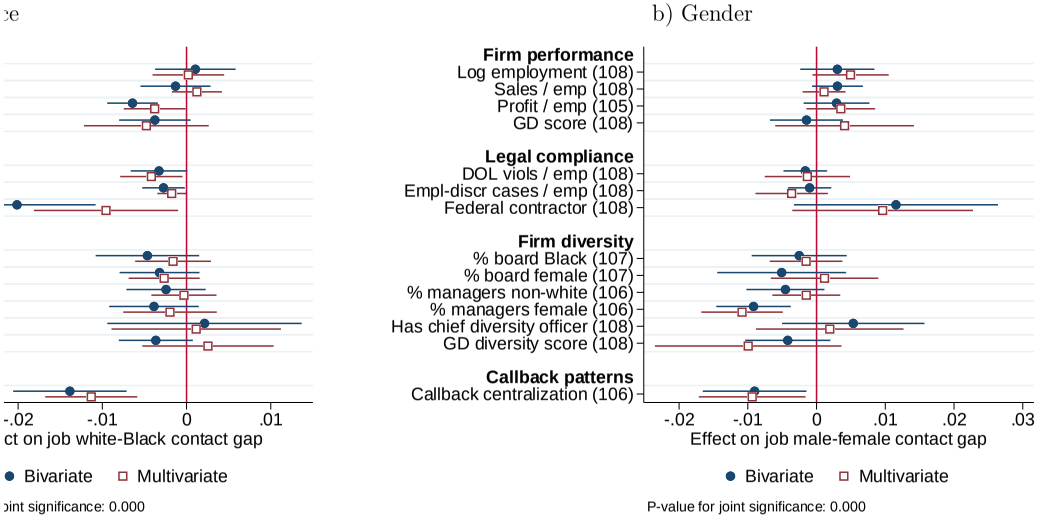


b) Gender



hip between establishment-level covariates and contact gaps for race and gender. Each relationship is estimated l contact gaps as the outcome. All jobs with defined contact gaps for each attribute and matched to the listed of jobs in each regression is in parentheses. Covariates are standardized to be mean zero, standard deviation 1 efficients from regressions of contact gaps on the covariate alone. "Firm FE" points include firm fixed effects. als based on robust standard errors. The omitted region category is the Northeast. Appendix C provides a initions and sources.

Figure 5: Relationships between contact gaps and firm characteristics

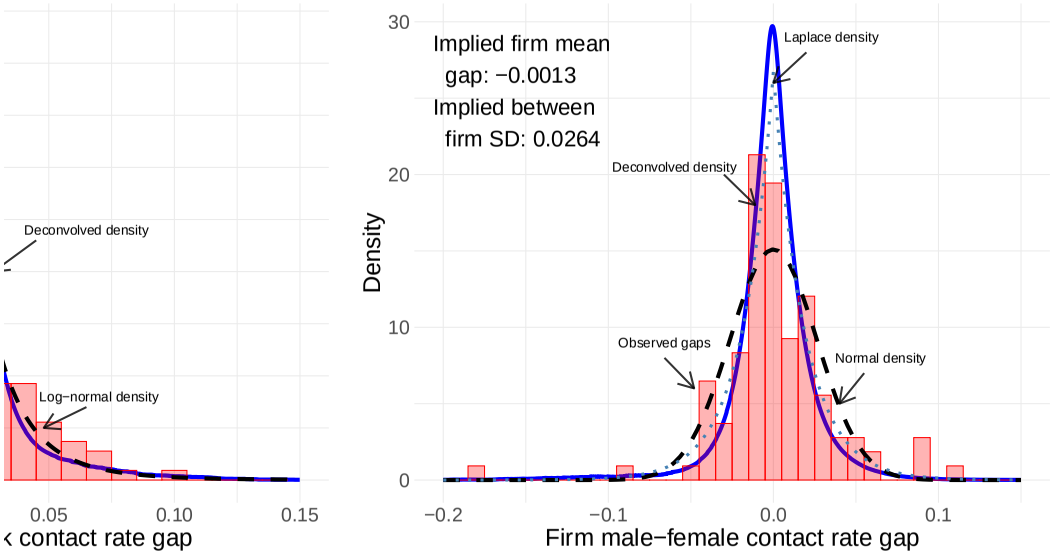


ship between firm-level covariates and contact gaps for race and gender. Each relationship is estimated by a ct gaps as the outcome, except for callback centralization which is estimated via split sample IV to account for outcome. The number of firms in each regression is in parentheses. Covariates are standardized to be mean , except for the binary indicators for federal contractor status and having a chief diversity officer. "Bivariate" ons of contact gaps on the covariate alone. "Multivariate" points plot effects when all covariates are included lization measured in half of the randomly split sample instrumented using its value in the other half and all ous regressors. Bars indicate 95% confidence intervals based on standard errors clustered at the firm level. cription of covariate definitions and sources.

ure 6: Deconvolution estimates of firm-level discrimination distributions

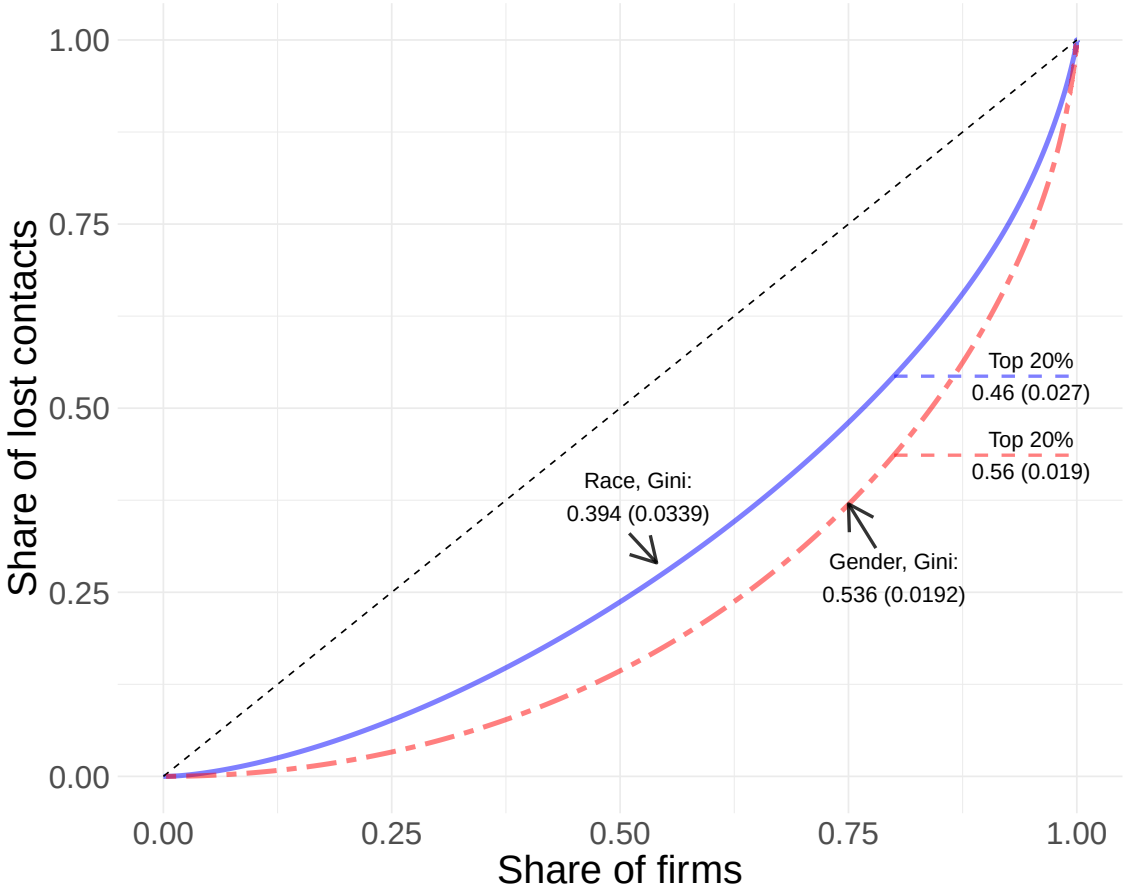
ce

b) Gender



metric estimates of the distribution of firm-specific contact gaps. Panel (a) presents estimates for white-Black) presents estimates for male-female differences. Red histograms show the distribution of estimated firm contact mates of population contact gap distributions. The population distributions are estimated by applying the nd Efron, 2020) to firm-specific z-score estimates, then numerically integrating over the empirical distribution stribution of contact gaps. The penalization parameter in the deconvolution step is calibrated so that the orresponding bias-corrected variance estimate from Table 4. In panel (a), the density of population z-scores Parametric comparison distributions shown with dashed and dotted lines are calibrated to have means and nvolved distributions.

Figure 7: Discrimination Lorenz curves



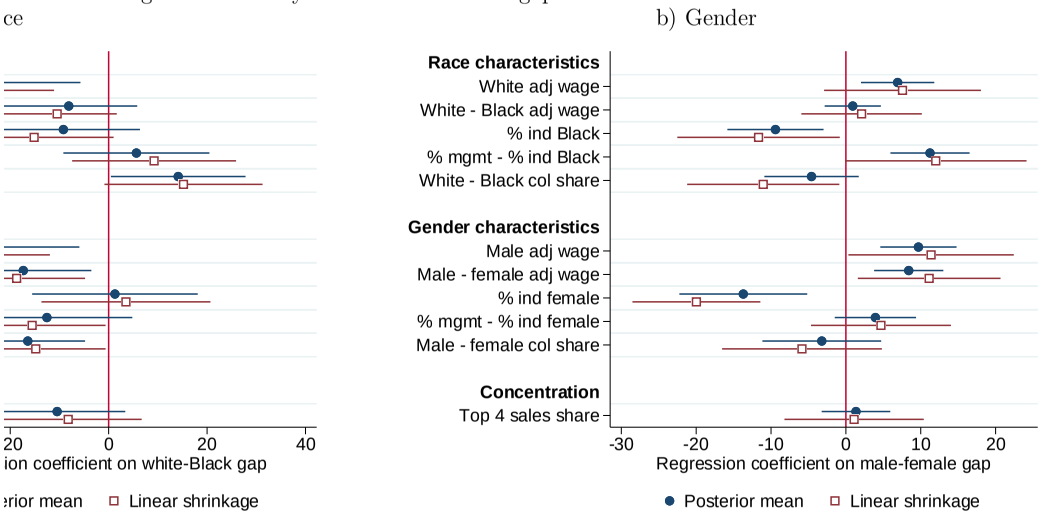
Notes: This figure displays Lorenz curves implied by the non-parametric deconvolution estimates of race and gender contact gap distributions in Figure 6. The solid blue curve is the Lorenz curve for the white/Black contact gap, and the dashed red curve is the Lorenz curve for the absolute value of the 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. Standard errors for Gini coefficients and top 20% shares are reported in parentheses. Standard errors are computed using 1,000 iterations of a parametric bootstrap redrawing a bootstrap observation $\hat{\Delta}_{fb}$ for each firm from a $\mathcal{N}(\hat{\Delta}_f, s_f^2)$ distribution.

Figure 8: Posterior means by industry



y-level averages of posterior mean contact gaps. The navy bars show average posteriors using deconvolved gaps as priors. The light red bars show averages of estimates shrunk linearly towards the grand mean with ratio $\hat{\theta}/(s_f^2 + \hat{\theta})$.

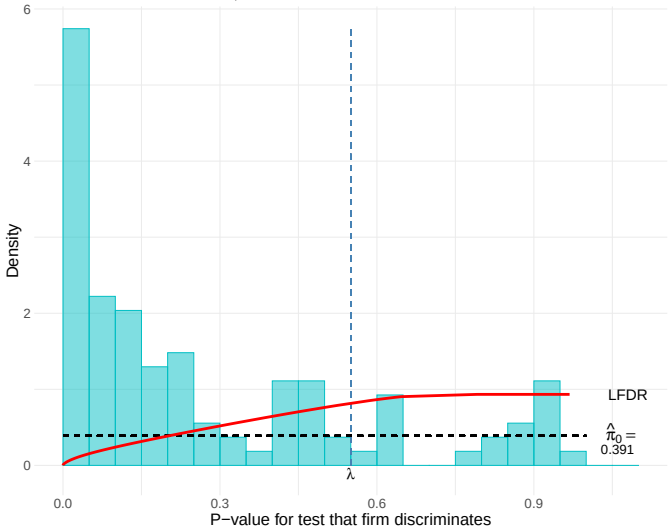
Figure 9: Industry correlates of contact gaps



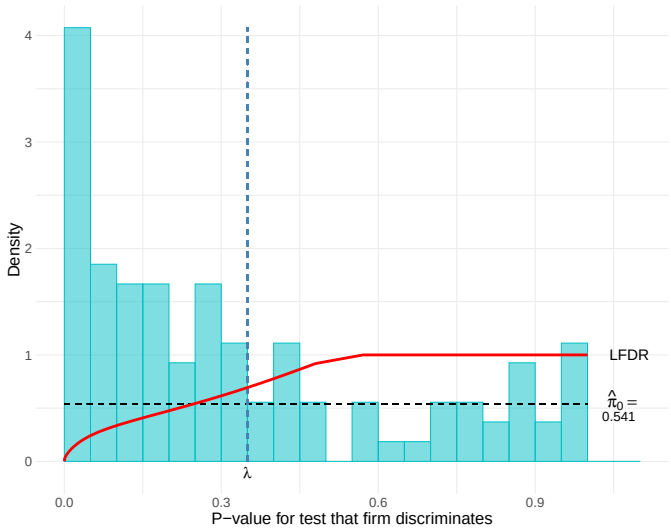
ns of industry characteristics on posterior mean contact gaps for race and gender. Points labeled "posterior r gaps formed using the distributions in Figure 6 as priors. Points labeled "linear shrinkage" show coefficients grand mean with weights given by the signal to noise ratio $\hat{\theta}/(s_f^2 + \hat{\theta})$. Outcomes are normalized to be mean . Appendix C provides a complete description of covariate definitions and sources.

Figure 10: *P*-value distributions and local false discovery rates

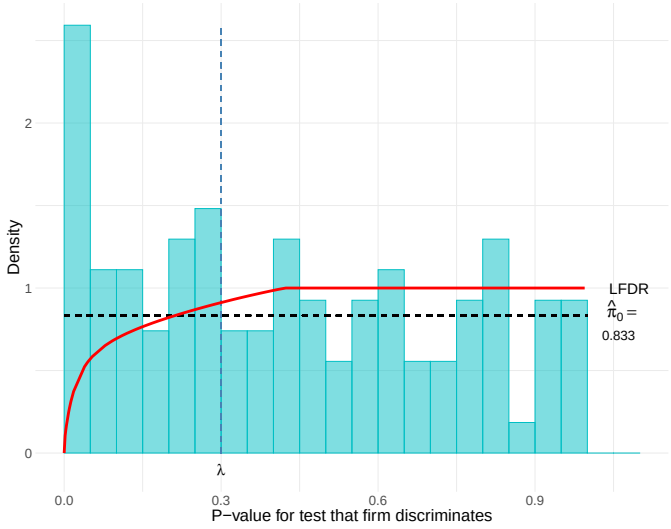
a) Race, one-sided



b) Race, two-sided

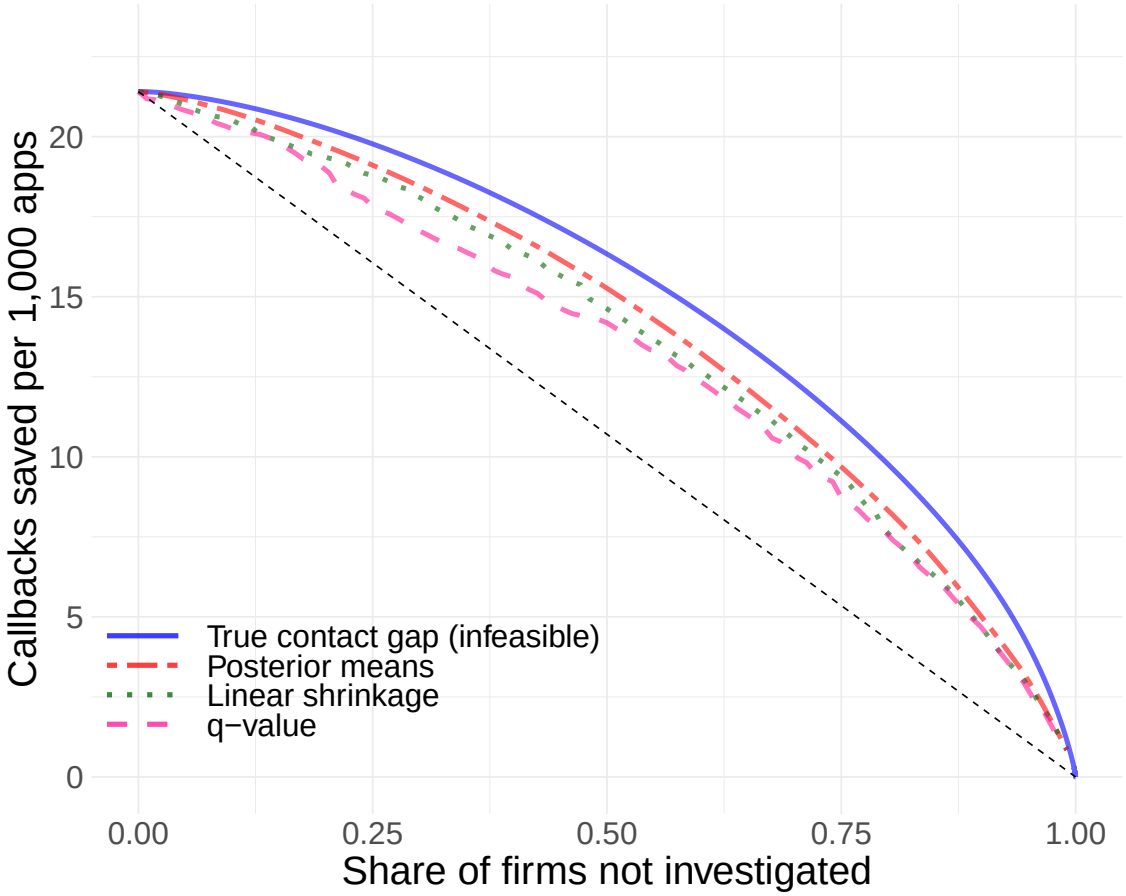


c) Gender, two-sided



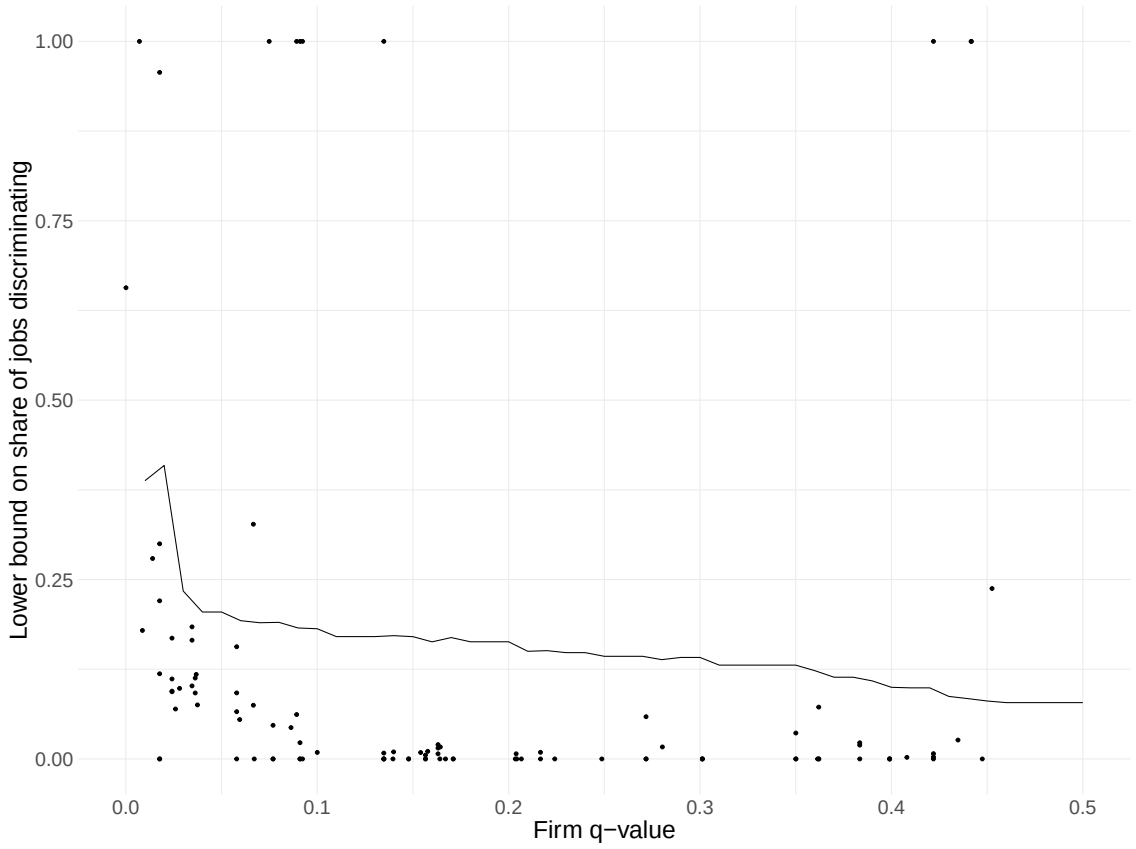
Notes: This figure plots distributions of p-values from firm-specific tests of the null hypothesis of no discrimination. Panel (a) shows results for one-sided tests of no discrimination against Black applicants, and panel (b) displays results for two-sided tests of equal contact rates for Black and white 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 11: 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 12: 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			3/4/14/15/72			

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) LPM	(2) Logit	(3) LPM	(4) 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	-0.0000389	-0.000671
	(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	83643	83643	65368	65368

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

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

	OLS			Logit		
	(1) White	(2) Black	(3) Difference	(4) White	(5) Black	(6) 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)
Associate 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	41837	41806	83643	41837	41806	83643
χ^2 stat for joint significance			14.71			14.54
p-value			0.0650			0.0687

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

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

			Contact gap SD		
	(1) χ^2 test of heterogeneity	(2) p-value for no discrim against:	(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	(0.0031)	(0.0032)	(0.0031)
Gender	205.2	M: 0.00	0.0267	0.0287	0.0269
	[0.000]	F: 0.05	(0.0038)	(0.0035)	(0.0038)
Over 40	144.6	Y: 0.22	0.0103	0.0044	0.0086
	[0.011]	O: 0.02	(0.0069)	(0.0158)	(0.0082)

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
	(0.0036)	(0.0018)	(0.0276)	(0.0152)	(0.0238)	(0.0131)
Std. dev.	0.1607	0.0186	0.9755	0.1155	0.7047	0.0837
	(0.0035)	(0.0035)	(0.0385)	(0.0360)	(0.0382)	(0.0341)
Corr. w/own slope	-0.4010	1.000	0.0519	1.000	0.0685	1
	(0.1098)	-	(0.2074)	-	(0.3092)	-
Corr. w/LPM slope	-0.4010	1.000	-0.4274	0.8944	-0.5045	0.8075
	(0.1098)	-	(0.1068)	(0.2095)	(0.1149)	(0.3074)
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	-	-
	(0.0034)		
	[0.038]	[0.668]	[0.583]
Industry	0.0141	0.0190	0.0048
	(0.0021)	(0.0029)	(0.0053)
	[0.000]	[0.000]	[0.112]
Job title	0.0136	0.0111	0.0034
SOC3 code	(0.0025)	(0.0043)	(0.0105)
	[0.000]	[0.007]	[0.527]
Hiring platform	0.0059	0.0024	0.0024
intermediary	(0.0025)	(0.0088)	(0.0071)
	[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
	0.0176	0.0150	0.0253	0.0255	0.0096	0.0088
ects	0.0003	-	-	0.0080	0.0004	-
	0.0000	0.0001	0.0000	0.0002	0.0000	0.0002
	11026	11026	10720	10720	10652	10652
	108	108	108	108	108	108
	51	47	51	47	51	47
1 firm	51	43	51	43	51	43
	0.0196	0.0196	0.0023	0.0023	0.0037	0.0037
	0.000	0.0008	0.000	0.000	0.071	.040
e effects	0.186	0.327	0.482	0.237	0.86	0.459

cted variance component estimates from two-way fixed effect models estimated using the leave-out procedure of olumns labeled "Job title" include fixed effects for the first three digits of each job's O*Net SOC code. Columns for the job's state. All variance and covariance estimates are job-weighted. Only jobs in the leave-job-out stimate. Dashes indicate negative variance estimates and hence undefined estimated standard deviations. "N mber of states or job titles in the connected set observed at 2 or more firms. The final two rows report p-values at all firm or job title / state fixed effects equal zero, computed using the heteroscedasticity robust procedure

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			
$\hat{\pi}_0$	0.370	0.455	0.808	0.802
# q-values <= 0.05	35	24	8	1
# q-values <= 0.1	55	36	10	1
λ	0.550	0.450	0.450	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			
$\hat{\pi}_0$	0.602	0.696	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).

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

	Federal contractor?	Contact gap	Std. err.	p-value	q-value	Posterior mean	Posterior 5th pctile	Posterior 95th pctile
rvices	Yes	0.0952	0.0197	0.0000	0.0001	0.0835	0.0450	0.1035
rvices	No	0.0507	0.0143	0.0003	0.0061	0.0354	0.0135	0.0673
rvices	No	0.0738	0.0220	0.0005	0.0073	0.0489	0.0192	0.0981
rvices	No	0.0787	0.0249	0.0010	0.0103	0.0498	0.0202	0.1031
es	No	0.0733	0.0250	0.0022	0.0158	0.0448	0.0187	0.0929
.	No	0.0469	0.0159	0.0020	0.0158	0.0286	0.0119	0.0595
.	Yes	0.0605	0.0219	0.0033	0.0176	0.0365	0.0154	0.0743
ıdise	Yes	0.0520	0.0187	0.0031	0.0176	0.0314	0.0132	0.0641
rvices	No	0.0613	0.0240	0.0060	0.0194	0.0370	0.0158	0.0725
.	No	0.0560	0.0214	0.0050	0.0194	0.0337	0.0143	0.0669
ng	No	0.0560	0.0222	0.0064	0.0194	0.0339	0.0144	0.0660
rvices	No	0.0540	0.0215	0.0068	0.0194	0.0327	0.0139	0.0634
	Yes	0.0511	0.0204	0.0069	0.0194	0.0310	0.0132	0.0599
ıdise	No	0.0427	0.0170	0.0068	0.0194	0.0259	0.0110	0.0502
res	Yes	0.0400	0.0159	0.0066	0.0194	0.0242	0.0103	0.0470
ırable	No	0.0386	0.0158	0.0080	0.0199	0.0235	0.0100	0.0450
turing	Yes	0.0350	0.0142	0.0078	0.0199	0.0213	0.0090	0.0409
ials	Yes	0.0373	0.0157	0.0093	0.0218	0.0229	0.0097	0.0433
es	Yes	0.0544	0.0240	0.0132	0.0292	0.0339	0.0143	0.0627
res	No	0.0400	0.0183	0.0152	0.0322	0.0252	0.0106	0.0460
ng	No	0.0340	0.0159	0.0172	0.0346	0.0217	0.0090	0.0392
ıdise	No	0.0423	0.0210	0.0229	0.0439	0.0277	0.0114	0.0494
estate	No	0.0278	0.0140	0.0257	0.0472	0.0183	0.0075	0.0325

of white-Black contact gaps for the 23 individual firms with q-values less than 0.05. P-values and q-values come thesis that the firm does not discriminate against Black applicants. To ensure that q-values are non-decreasing low 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 ans and percentiles are empirical Bayes posteriors constructed using the estimated distribution in Figure 6 as

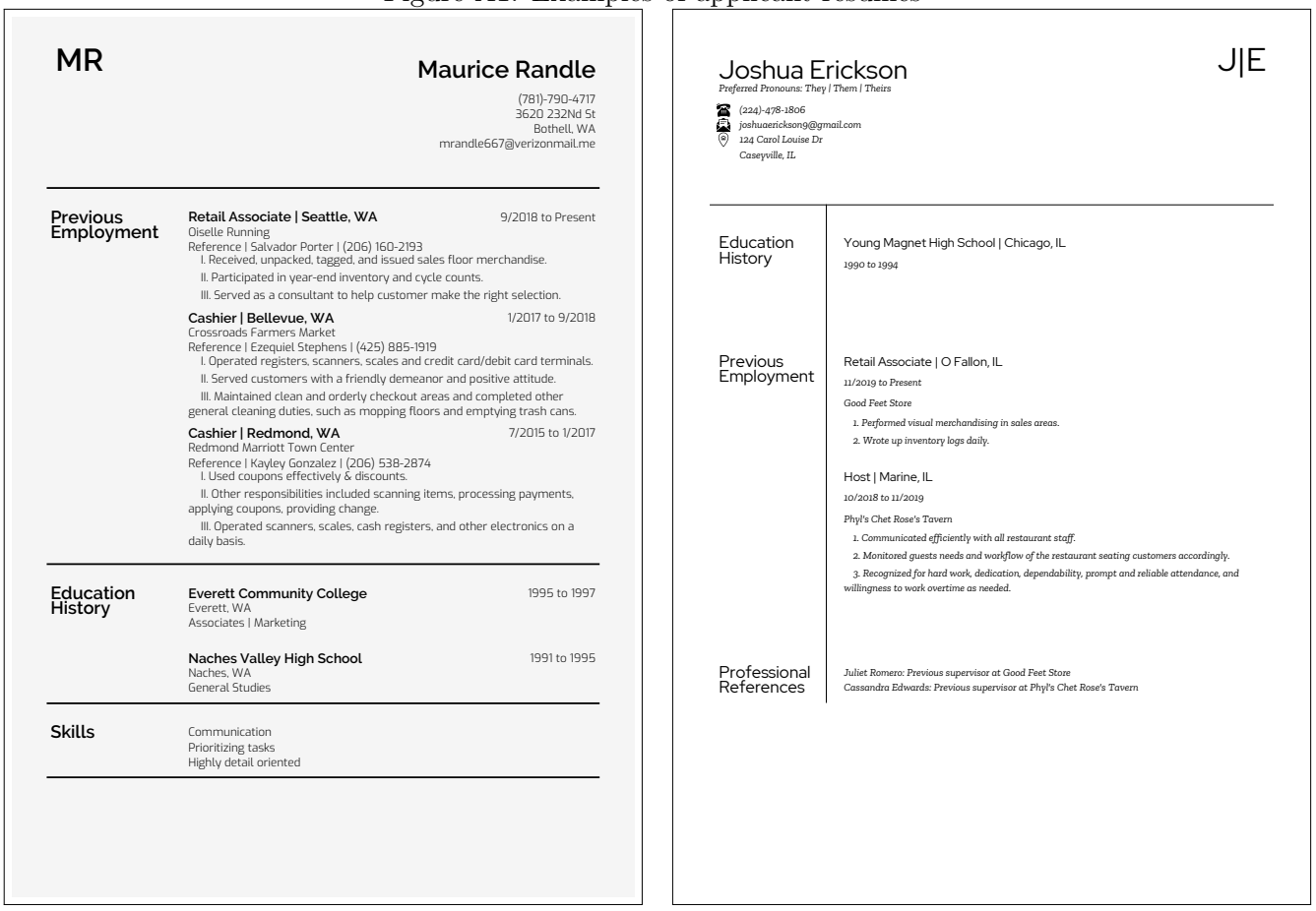
Table 10: Discrimination estimates and detection by industry

SIC	Industry	N firms	Race			Gender		
			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	0.900	-0.004	0	0.993
23	Apparel manufacturing	2	0.021	1	0.170	0.007	0	0.702
24-35	Other manufacturing	4	0.018	0	0.361	0.012	0	0.669
42-47	Freight / transport	4	0.011	0	0.822	0.001	0	0.941
48	Communications	2	0.017	0	0.340	0.013	0	0.972
49	Electric / gas	3	0.015	0	0.339	0.002	0	0.980
50	Wholesale durable	2	0.017	0	0.293	0.034	0	0.555
51	Wholesale nondurable	11	0.018	1	0.456	0.005	0	0.865
52	Building materials	3	0.014	1	0.544	0.012	0	0.849
53	General merchandise	12	0.023	3	0.276	-0.001	0	0.867
54	Food stores	5	0.025	1	0.356	0.009	0	0.821
55	Auto dealers / services	8	0.040	6	0.127	0.005	0	0.882
56	Apparel stores	4	0.025	1	0.253	-0.061	1	0.416
57	Furnishing stores	4	0.022	2	0.304	-0.006	0	0.787
58	Eating/drinking	5	0.027	2	0.303	0.003	0	0.926
59	Other retail	7	0.022	3	0.314	-0.002	0	0.971
60-61	Banks / credit	5	0.010	0	0.651	0.002	0	0.778
62	Securities brokers	2	0.010	0	0.410	-0.011	0	0.654
63-65	Insurance / real estate	8	0.013	1	0.463	-0.003	0	0.915
70	Accommodation	2	0.015	0	0.527	0.001	0	1.000
73	Business services	3	0.012	0	0.539	0.000	0	0.942
75-76	Auto / repair services	3	0.013	0	0.474	0.015	0	0.624
80	Health services	5	0.016	1	0.726	-0.009	0	0.909
87	Engineering services	2	0.009	0	0.348	-0.001	0	0.965

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 10. The distribution of gender LFDRs is depicted in Panel (c) of Figure 10.

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
joshuaericksn9@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.*

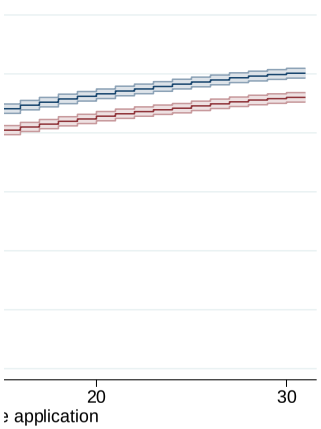
Professional References

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

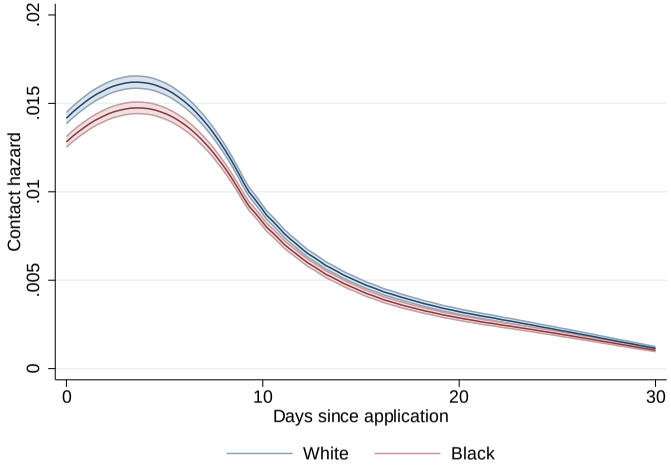
Notes: This figure presents two examples of randomly generated resumes used in the experiment. Resumes are formatted using a combination of pre-set options specifying length, fonts, text sizes, section header names, and layouts, with controls to ensure resumes that overflow one page are not generated.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

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

robability

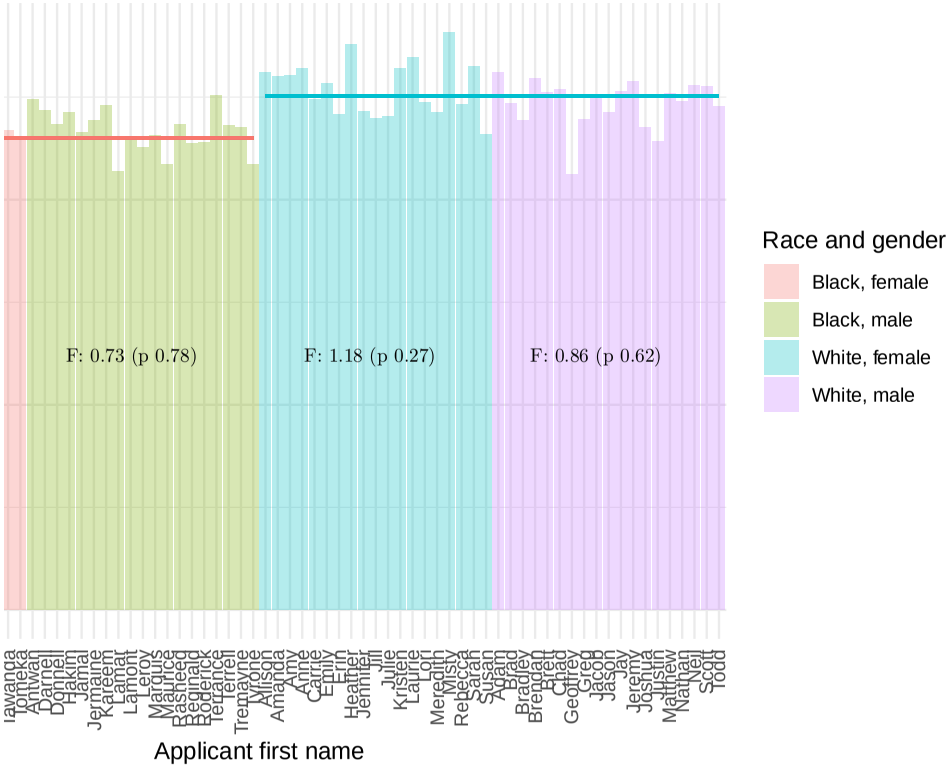


b) Smoothed contact hazard



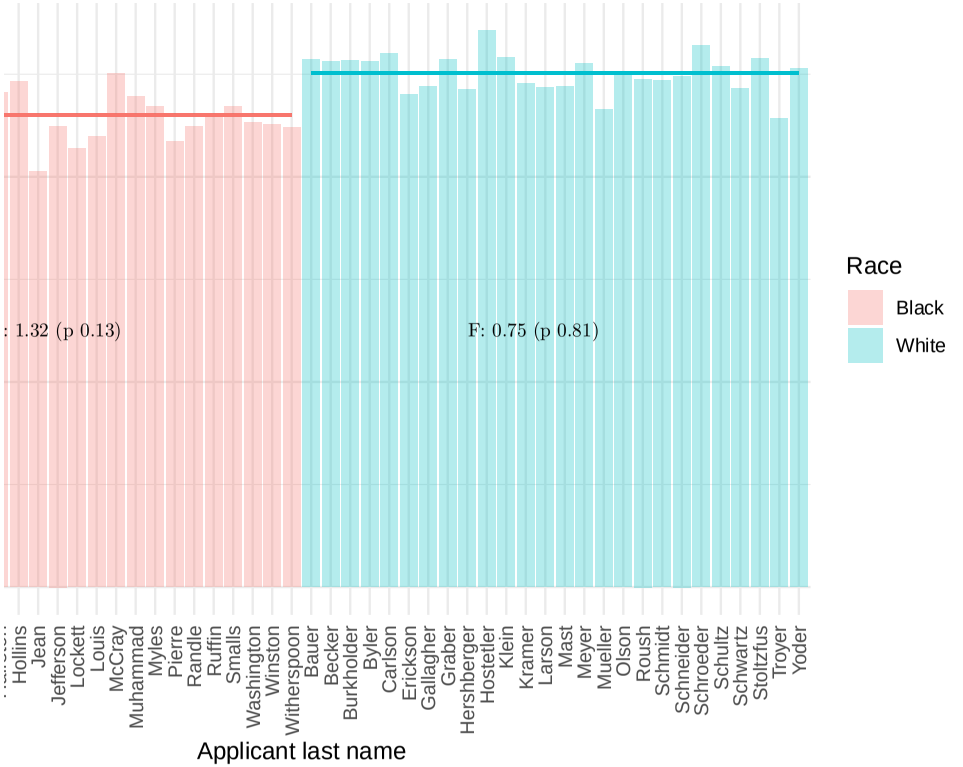
babilities and hazards as functions of days since application. Contact probabilities correspond to Kaplan-Meier are Kaplan-Meier hazard estimate smoothed using the Epanechnikov kernel. Shaded areas represent pointwise

Figure A3: Callbacks by applicant first name



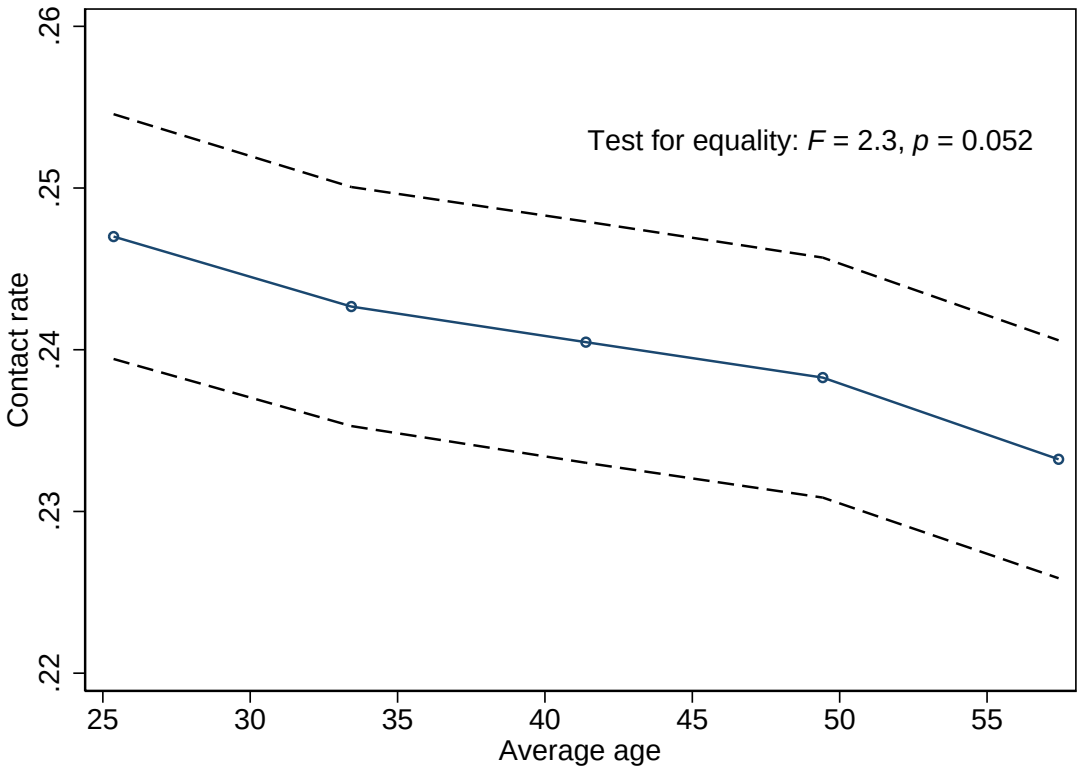
act rates by applicant first name, organized by race and gender group. The horizontal bars show race group
alues come from joint tests of the hypothesis that contact rates are equal across names separately by race and

Figure A4: Callbacks by applicant last name



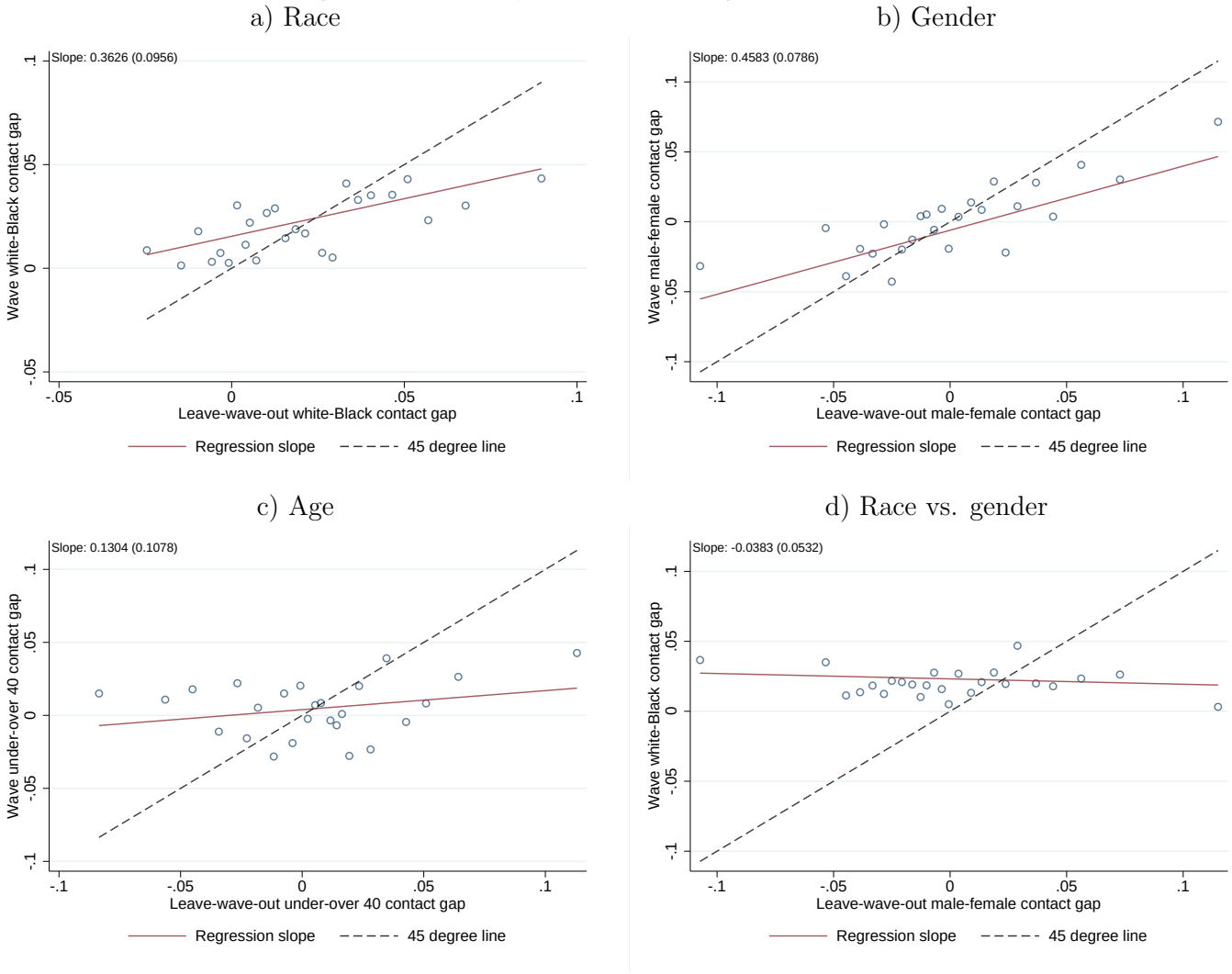
ıct rates by applicant last name, organized by race. The horizontal bars show race group mean contact rates.
: tests of the hypothesis that contact rates are equal across names separately by race.

Figure A5: Contact rates by age category



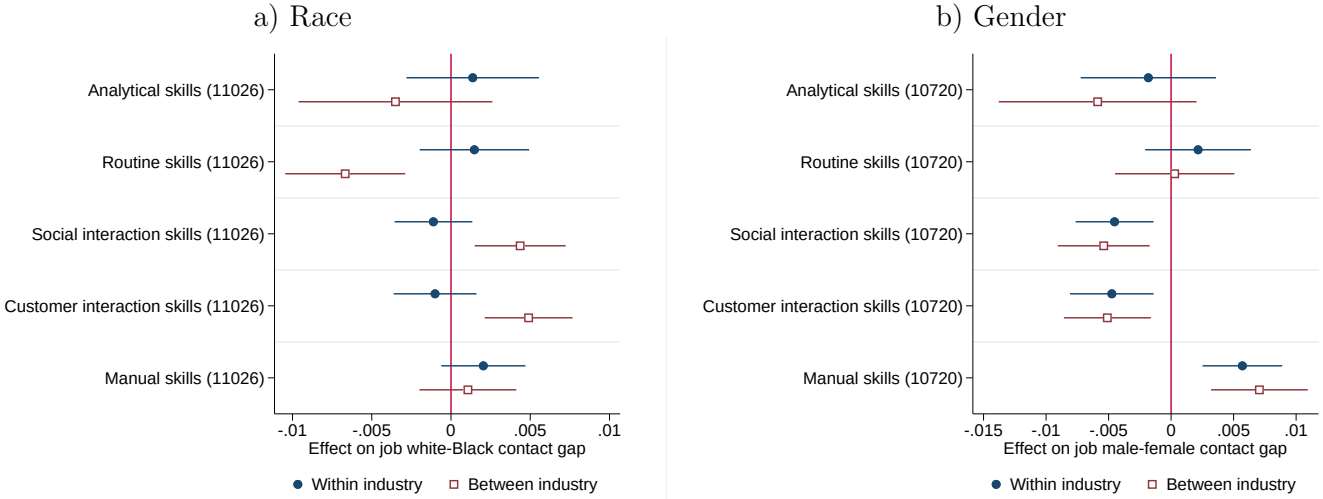
Notes: This figure plots average 30-day contact rates by quintile of applicant age at the time of application. Estimates come from regressions of a contact indicator on indicators for age quintile, controlling for wave indicators. The horizontal axis plots average age in each quintile. The vertical axis 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: Stability of firm contact gaps across waves



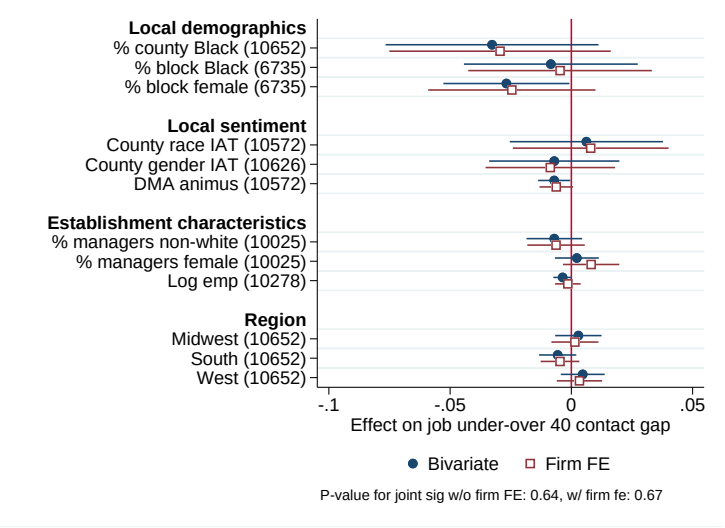
Notes: This figure presents binned scatter plots of firm-specific wave-average contact gaps vs. leave-wave-out firm-specific average contact gaps. Panel (a) reports results for the white/Black difference in contact rates. Panel (b) shows results for the male/female difference in contact rates. Panel (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 vingtiles 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 A7: Within and between industry relationship between contact gaps and task content



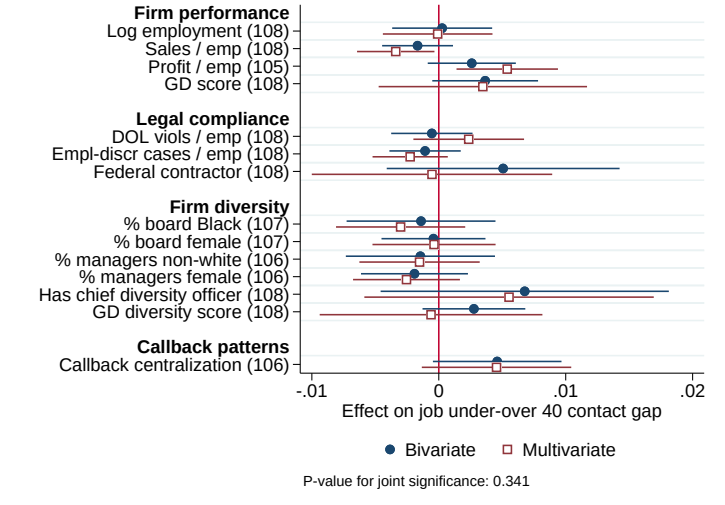
Notes: This figure plots the relationship between O*Net measures of job-level task content and contact gaps for race and gender within and between industry. The within relationship is estimated with a linear regression with job-level contact gaps as the outcome and two-digit industry fixed effects. The between relationship is estimated by instrumenting job task content with industry dummies. 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. Bars indicate 95% confidence intervals based on robust standard errors. Appendix C provides a complete description of task definitions and sources.

Figure A8: Relationships between age contact gaps and establishment characteristics



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 A9: Relationships between age contact gaps and firm characteristics

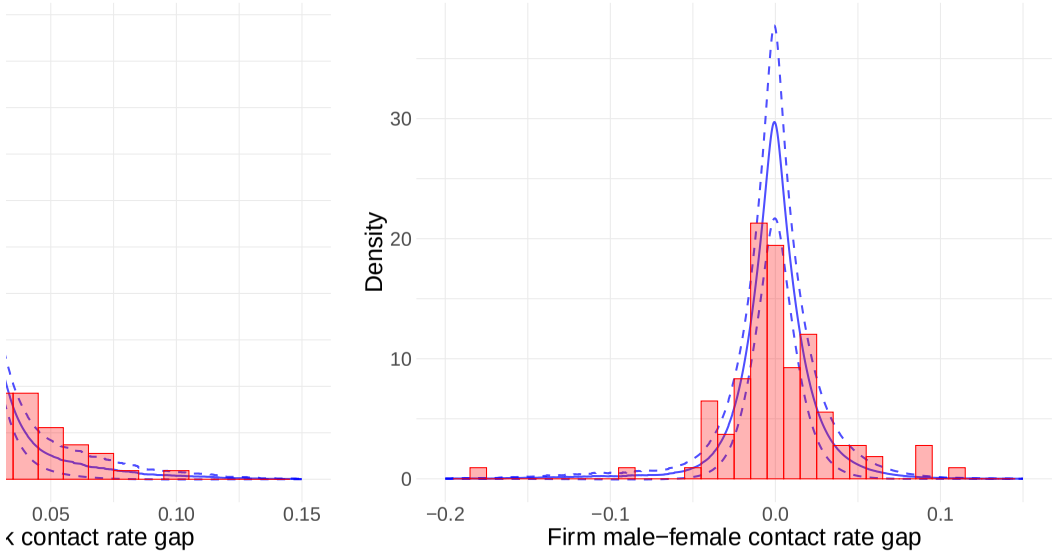


Notes: This figure plots relationships between firm-level covariates and contact gaps for application age under vs. over 40. Each relationship is estimated with a linear regression with job-level contact gaps as the outcome. All covariates are standardized to be mean zero, standard deviation 1 in sample."Bivariate" 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.

e A10: Confidence intervals on deconvolutions of firm-level discrimination

ce

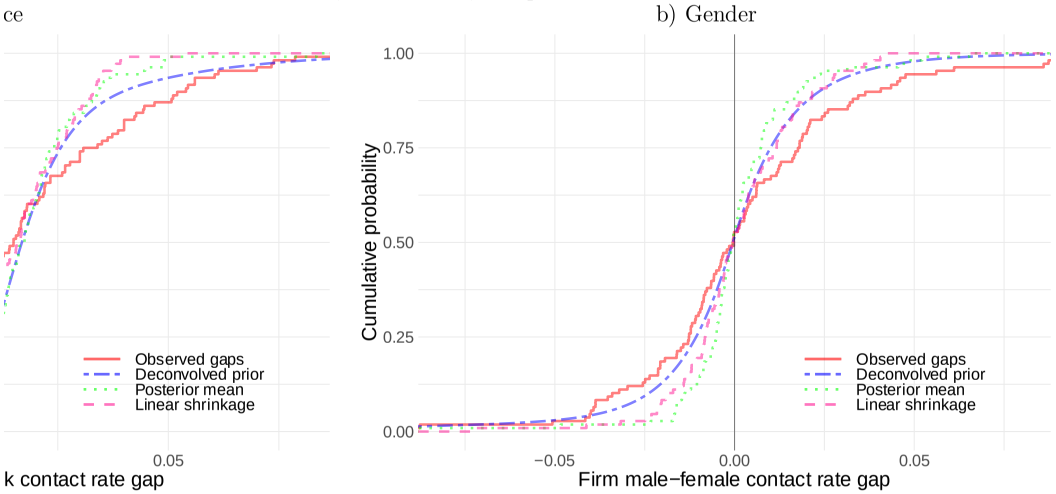
b) Gender



metric estimates of the distribution of firm-specific contact gaps and point-wise 95% confidence intervals. Panel
k contact rate differences, and panel (b) presents estimates for male-female differences. Red histograms show
ıtact gaps. Blue lines shows estimates of population contact gap distributions. The population distributions are
eR package (Narasimhan and Efron, 2020) to firm-specific z-score estimates, then numerically integrating over
d errors to recover the distribution of contact gaps. Since the estimated density of Δ is a linear combination
rd errors can be computed using the delta method and the variance-covariance of $\hat{g}_\mu$ produced by Narasimhan
parameter in the deconvolution step is calibrated so that the resulting distribution matches the corresponding
ı Table 4. In panel (a), the density of population z-scores is constrained to be weakly positive.

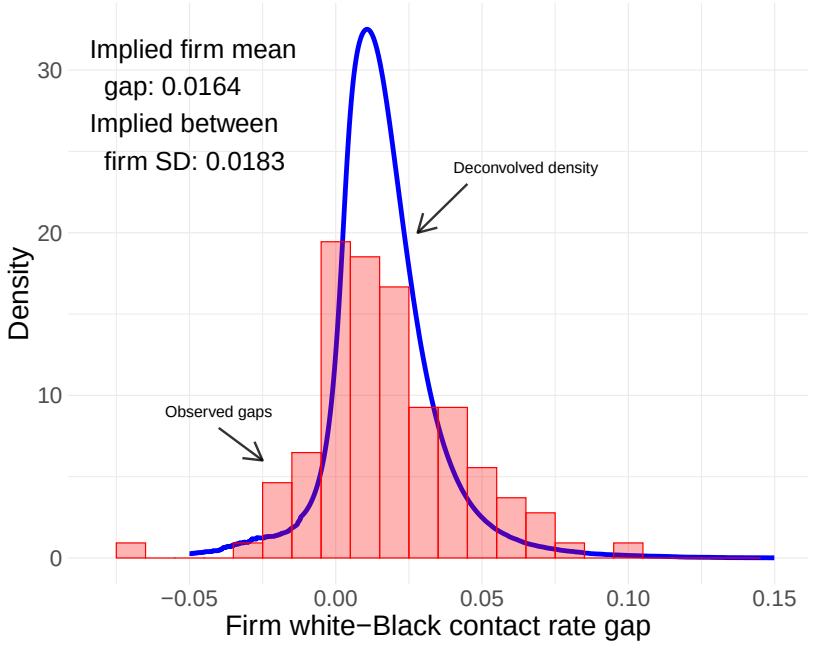
ure A11: Distribution of observed, deconvolved, and posterior estimates

ce



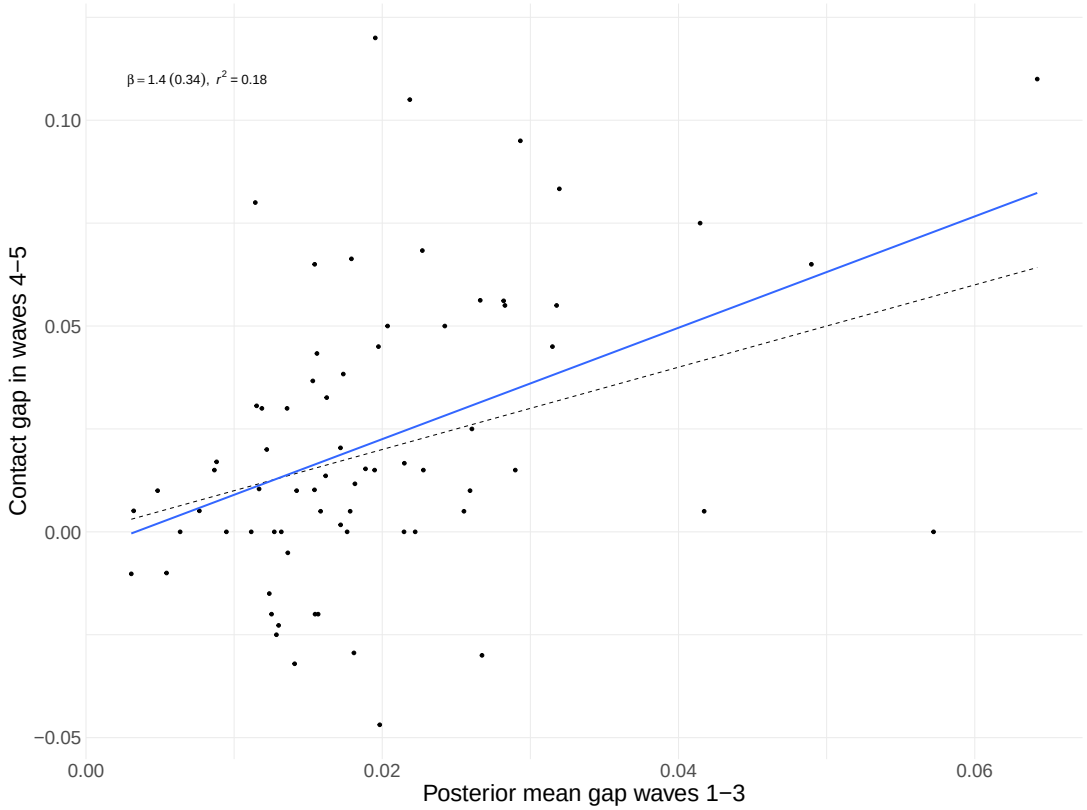
s of the distribution of firm-specific contact gaps for race and gender. The red solid line presents the empirical ashed line plots the CDF of population contact gaps based on the deconvolution estimates in Figure 6. The CDF of posterior means, constructed treating the deconvolved density as prior knowledge. The pink dashed nates shrunk linearly towards the grand mean with weights given by the signal-to-noise ratio $\hat{\theta}/(s_f^2+\hat{\theta})$, where e variance of contact gaps across firms.

Figure A12: Deconvolution of firm-level racial discrimination without support restriction



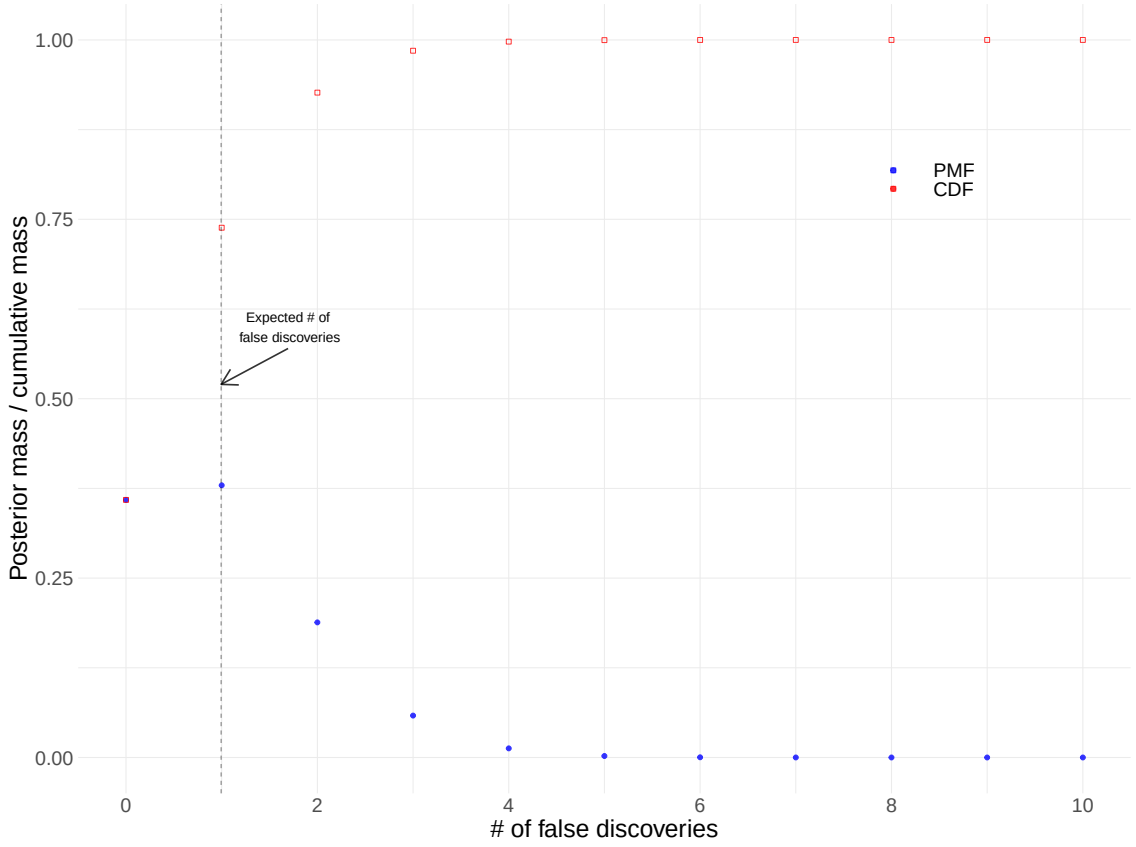
Notes: This figure presents non-parametric estimates of the distribution of firm-specific white-Black contact rate differences. The red histogram shows the distribution of estimated firm contact gaps. Blue line shows estimates of the population contact gap distributions. The population distributions are estimated by applying the deconvolveR package (Narasimhan and Efron, 2020) to firm-specific z-score estimates, then numerically integrating over the empirical distribution of standard errors 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 A13: Out-of-sample predictive power of racial contact gap posteriors



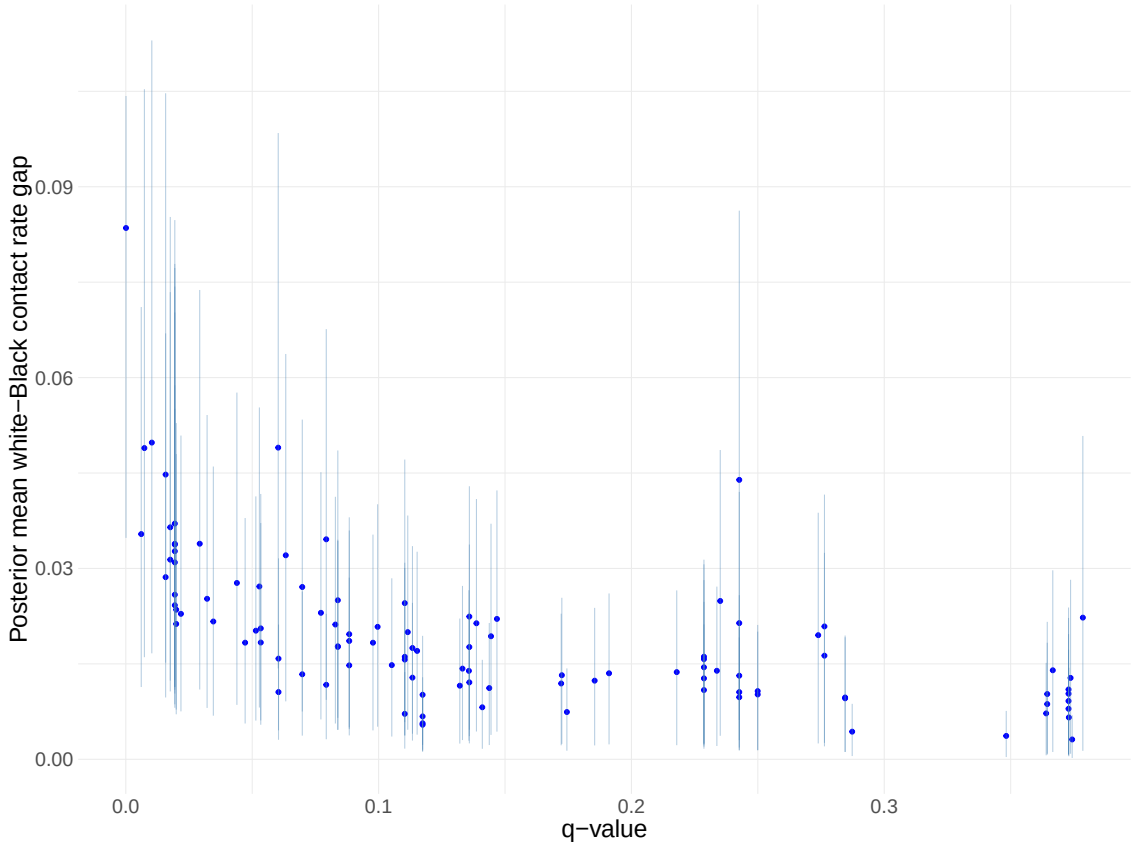
Notes: This figure plots posterior mean white-Black contact gaps computed using data from waves 1-3 against observed gaps in waves 4-5 for the sample of firms included in all five waves. Posterior means are computed using the population contact gap distributions estimated by applying the deconvolveR package (Narasimhan and Efron, 2020) to firm-specific z-score estimates from waves 1-3. The penalization parameter in the deconvolution step is calibrated so that the resulting distribution matches the corresponding bias-corrected variance estimate. The black dotted line is the 45 degree line. The blue line is the least squares fit. Adjusting for the noise in the wave 4 and 5 estimates yields a bias corrected R^2 of 0.5, or a correlation between predictions in later waves and the latent true contact gaps of $\sqrt{0.5} = 0.71$.

Figure A14: Posterior false discovery distribution among 23 firms with low q-values



Notes: This figure plots EB posterior estimates of the probability mass function and cumulative distribution of false discoveries among the 23 firms with q-values below 0.05 for race. Posterior was calculated using the Poisson binomial distribution implied by the 23 firms' LFDR estimates plotted in Figure 10. The dotted line denotes the expected number of false discoveries among these firms, which coincides with the mean of their LFDRs.

Figure A15: 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) χ^2 test of heterogeneity	(2) p-value for no discrim against:	(3) Bias-corrected	(4) Cross-wave	(5) Cross-state
Race	229.5	W: 1.00	0.0184	0.0171	0.0182
	[0.000]	B: 0.00	(0.0029)	(0.0032)	(0.0031)
Gender	124.2	M: 0.06	0.0207	0.0213	0.0200
	[0.000]	F: 0.03	(0.0044)	(0.0043)	(0.0045)
Over 40	90.2	Y: 0.15	0.0098	0.0096	0.0099
	[0.072]	O: 0.02	(0.0060)	(0.0067)	(0.0057)

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 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 2 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). Column 3 reports bias-corrected estimates of standard deviations of average contact gaps across firms based on covariances between job-level contact gaps. 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 the Appendix. Standard errors for all variance estimators are produced by job-clustered weighted bootstrap.

Table A2: Variance components for other resume attributes

			Contact gap SD		
	(1) χ^2 test of heterogeneity	(2) p-value for no discrim against:	(3) Bias-corrected	(4) Cross-wave	(5) Cross-state
LGBTQ Club Member	88.0	W: 1.00	-	-	-
	[0.885]	B: 0.98			
Gender Neutral Pronouns	126.5	Y: 0.92	0.0198	0.0177	0.0147
	[0.076]	O: 0.65	(0.0156)	(0.0176)	(0.0208)

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 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 2 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). Column 3 reports bias-corrected estimates of standard deviations of average contact gaps across firms based on covariances between job-level contact gaps. 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 the Appendix. Standard errors for all variance estimators are produced by job-clustered weighted bootstrap. 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	0.070	0.090	0.026
	(0.000)	(0.000)	(0.000)
Prevalence bound	0.073	0.000	0.014
	(0.012)	(0.001)	(0.021)

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 viols/emp: Includes all wage and hour compliance violations 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 adj wage, white - Black adj wage, male adj wage, male - female adj wage: Constructed using the CPS Monthly Outgoing Rotation Groups from 2009 to 2019,

Table C1: Summary statistics for firm-level covariates

	Mean	SD	Median
Firm performance			
Log employment	11.067	1.01	10.922
Sales / emp ($M)	0.331	0.36	0.238
Profit / emp ($M)	0.101	0.08	0.078
GD score	3.566	0.32	3.600
Legal compliance			
DOL viols / emp	0.136	0.37	0.002
Empl-discr cases / thousand emp	0.048	0.13	0.020
Federal contractor	0.667	0.47	1.000
Firm diversity			
% board Black	0.088	0.07	0.091
% board female	0.257	0.10	0.255
% managers non-white	0.257	0.09	0.250
% managers female	0.493	0.39	0.449
Has chief diversity officer	0.167	0.37	0.000
GD diversity score	3.816	0.33	3.800
Callback patterns			
Callback centralization	-1.117	0.38	-1.073
Observations	108		

Notes: This table reports summary statistics for firm-level covariates. See Appendix C for full details on the sources and construction of each variable.

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 correspond to the same coefficients from regressions with an indicator for female (or Black when constructing gender gaps), education (6 categories), and a quartic in age also included. 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.

Table C2: Firm-level predictors of centralization

Firm performance	
Log employment	-0.144
	(0.116)
Sales / emp	-0.0682
	(0.0813)
Profit / emp	-0.00380
	(0.0797)
GD score	-0.211
	(0.173)
Legal compliance	
DOL viols / emp	-0.0836
	(0.121)
Empl-discr cases / emp	0.0576
	(0.0427)
Federal contractor	0.559**
	(0.265)
Firm diversity	
% board Black	0.178
	(0.116)
% board female	-0.0153
	(0.0982)
% managers non-white	0.00869
	(0.123)
% managers female	0.0553
	(0.0986)
Has chief diversity officer	0.148
	(0.211)
GD diversity score	0.223
	(0.157)
Observations	10500

Notes: This table reports the multivariate relationship between centralization and other firm-level predictors. All predictors except the binary indicators for federal contractor status and having a chief diversity officer are normalized to have standard deviation of 1. As with firm-level relationships reported in Figure 5, the regression is estimated on job-level data with firm-clustered standard errors. See Appendix C for full details on the sources and construction of each variable.

- 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").
  - 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

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

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, 1988; Aaronson, Barrow and Sander, 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 $(\tau_1,\tau_2) \in \{1,...,T_f\}^2$

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

where $\hat{\bar{\Delta}}_{f\tau}$ 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

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

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:

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

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:

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

Appendix E Alternative Deconvolution Estimates

This section explores the robustness of estimated population contact gap distributions to alternative models for the relationship between estimated gaps, $\hat{\Delta}_f$, and their standard errors, s_f. Our baseline analysis assumes that s_f is independent of the population z-score Δ_f/s_f. After applying the Efron (2016) deconvolution estimator to the sample z-scores $\hat{\Delta}_f/s_f$, we recover the distribution of Δ_f by numerically integrating against the empirical distribution of standard errors. Here we consider three alternatives: a variance stabilizating transformation approach, a "local deconvolution" approach that separately estimates population distributions among groups of firms with similar standard errors, and a non-parametric approach that estimates the joint distribution of contact gaps and standard errors.

Appendix E.1 Variance stabilizing transformation

If one assumes a parametric model for the dependence of the firm specific variances on the latent contact gaps of the form $s_f^2 = h(\Delta_f)$, then heteroscedasticity can be eliminated via the variance stabilizing transformation:

$$y(t) = \int_{\infty}^{t} h(x)^{-1/2} dx.$$

Note that $\frac{d}{dt}y(t) = h(t)^{-1/2}$. Hence, standard delta-method reasoning implies that $y(\hat{\Delta}_f)|\Delta_f \sim \mathcal{N}(y(\Delta_f), 1)$. Applying the deconvolution estimator of Efron (2016) to the transformed estimates $y(\hat{\Delta}_f)$, one can then generate an estimate of the population distribution of Δ_f using the change of variables $\hat{g}_\Delta(x) = \hat{g}_{y(\Delta)}(x)h'(x)$, where $\hat{g}_t(\cdot)$ is the estimated density of t.

To implement this approach, we allow for non-linear dependence of the (squared) standard errors on contact gaps by assuming that

$$h(\Delta) = \alpha + \beta_1\Delta + \beta_2\Delta^2 \quad \text{for } \Delta \in \mathcal{S},$$

where $\mathcal{S}$ is the support of the population contact gap under study. We use split-sample IV (Angrist and Krueger, 1995) to estimate the parameters (β_1, β_2). Splitting each firm's jobs into two groups $g \in \{0, 1\}$, we proxy each group's values of Δ_f and Δ_f^2 with $\hat{\Delta}_{fg}$ and $\hat{\Delta}_{fg}^2 - s_{fg}^2$, respectively, where s_{fg} is the standard error of $\hat{\Delta}_{fg}$. We then estimate β_1 and β_2 via a regression of s_{fg}^2 on $(\hat{\Delta}_{fg}, \hat{\Delta}_{fg}^2 - s_{fg}^2)$ using as instruments $(\hat{\Delta}_{f(g)}, \hat{\Delta}_{f(g)}^2 - s_{f(g)}^2)$, where $(g) = 1 - g$ refers to the omitted group. To minimize uncertainty attributable to the splitting process, we take the median across 1,000 iterations of this procedure.

Figure E1 presents the resulting deconvolved population distributions of contact gaps for race and gender. As in the primary estimates, the race gap distribution exhibits a

peak close to zero and a fat right tail of heavy discriminators. The distribution of gender gaps continues to show concentrated mass near zero and severe discriminators in both tails. Figure E2 summarizes the concentration of discrimination based on the variance-stabilized approach by plotting the Lorenz curves implied by the deconvolved density $\hat{g}_\Delta$. Similar to our baseline estimates in Figure 7, these curves imply discrimination is concentrated in a relatively small share of firms for both race and gender. Finally, Figure E3 shows the estimated racial contact gap distribution based on the variance-stabilization approach without restricting the density of Δ_f to be weakly positive, which produces minimal changes to the results.

Appendix E.2 Local deconvolutions

A less parametric approach to dealing with heteroscedasticity in the contact gaps is to simply estimate the deconvolution within bins defined by ranges of s_f. This approach weakens the requirement that s_f be independent of the population z-score Δ_f/s_f in the full population to a requirement that independence only hold among firms with similar s_f. To implement this approach, we split firms into two groups $k \in \{H, L\}$ by whether their contact gap standard error falls above / below the sample median standard error. We then apply the deconvolution estimator of Efron (2016) to the sample z-scores in each group, $\hat{\Delta}_{fk}/s_{fk}$, and recover the group-specific population contact gap density, $g_{\Delta,k}$, by integrating against the empirical distribution of standard errors in that group. The marginal density of contact gaps is then given by the mixture:

$$\hat{g}_\Delta(x) = \frac{1}{2} g_{\Delta,H} + \frac{1}{2} g_{\Delta,L}$$

We use a common penalization parameter in the deconvolution step for both groups and calibrate it so that the resulting marginal distribution matches the corresponding bias-corrected variance estimate from Table 4.

Figure E4 shows the resulting group-specific densities for both race and gender. Figure E5 shows the corresponding marginal densities. As in the primary estimates, the race density shows concentrated mass close to zero and fat right tail. The gender density is strongly peaked at zero. Both densities continue to show that discrimination is strongly concentrated in a relatively small share of firms, as shown in Lorenz curves presented in Figure E6.

The close agreement of the top 20% share estimates found in Figures 7, E6, and E2 is reassuring and suggests our modeling of heteroscedasticity patterns is not an important factor driving our concentration results.

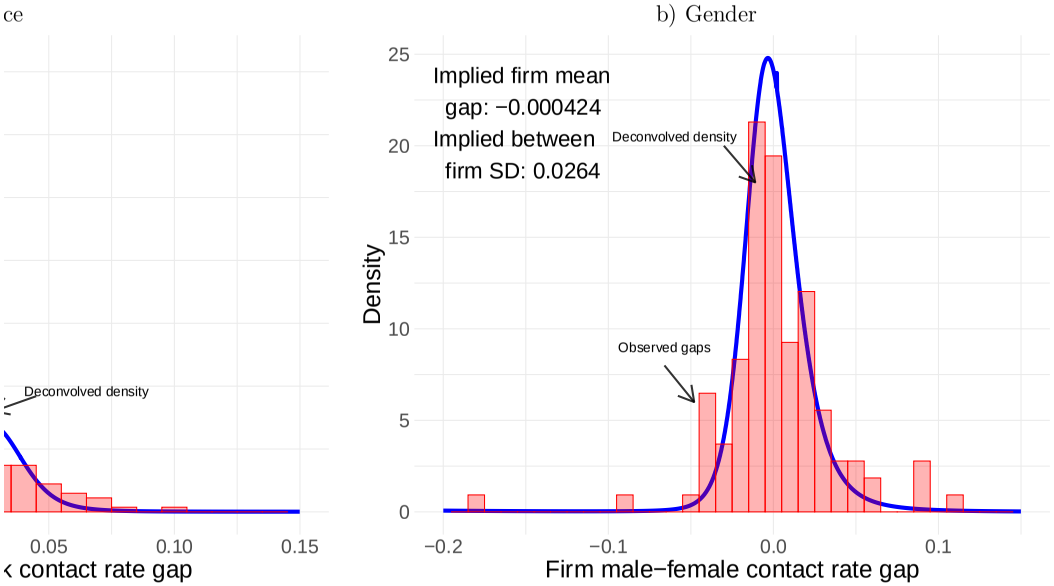
Appendix E.3 NPMLE

As a final approach to accounting for heteroscedasticity, we estimate a non-parametric mixing distribution over contact gaps and standard errors that allows for unrestricted dependence between these objects. To implement this approach, we use the approximation to the Kiefer-Wolfowitz non-parametric maximum likelihood estimator (NPMLE) developed by Koenker and Mizera (2014) and implemented in the REBayes package of Koenker and Gu (2017).

Figure E7 presents the resulting discrete marginal densities of contact gaps for race and gender. NPMLE estimates of the distribution of racial discrimination show similar patterns to our earlier spline approximations, with a concentrated mass of firms exhibiting limited discrimination and a fat tail of more heavy discriminators. Gender estimates show substantial mass near zero and smaller mass points in the extremes of both tails.

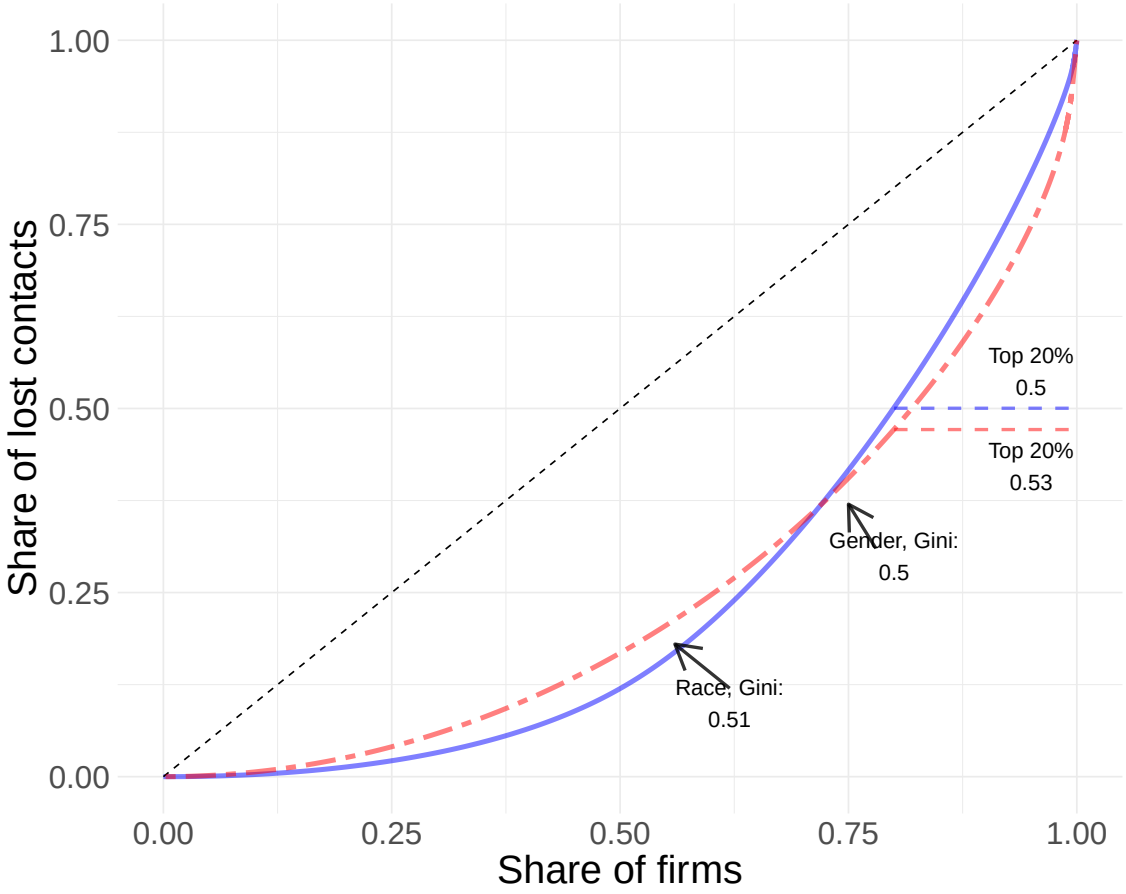
Figure E8 shows that these densities also imply substantial concentration of discrimination among a subset of employers. For comparability with prior results, we linearly interpolate between mass points, which yields kinks in the Lorenz curves. The interpolated top 20% shares are slightly higher than in our baseline specification utilizing spline approximations, suggesting again that our concentration finding is highly robust. The Gini coefficients are also close to those found in our baseline specification, with the race Gini slightly higher and the gender Gini slightly lower than the corresponding estimates in Figure 7.

1: Variance-stabilized deconvolution of firm-level discrimination distributions



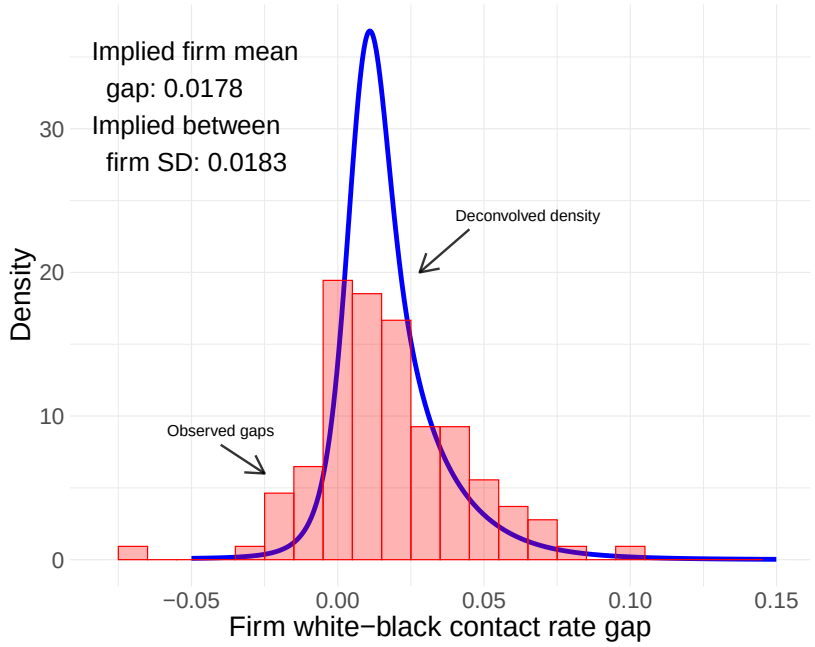
metric estimates of the distribution of firm-specific contact gaps. Panel (a) presents estimates for white-Black) presents estimates for male-female differences. Red histograms show the distribution of estimated firm contact opulation contact gap distributions. The population distributions are estimated by applying the deconvolveR 20) to variance-stabilized estimates of firm-specific contact gaps. The variance-stabilizing transformation is $+\beta_1\Delta_f+\beta_2\Delta_f^2$, with α, β_1, and β_2 estimated via split-sample IV (Angrist and Krueger, 1995). The estimated ed gaps is transformed into the distribution of Δ_f using the change of variables formula. The penalization is calibrated so that the resulting distribution matches the corresponding bias-corrected variance estimate from population Δ_f is constrained to be weakly positive.

Figure E2: Variance-stabilized 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 E1. 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.

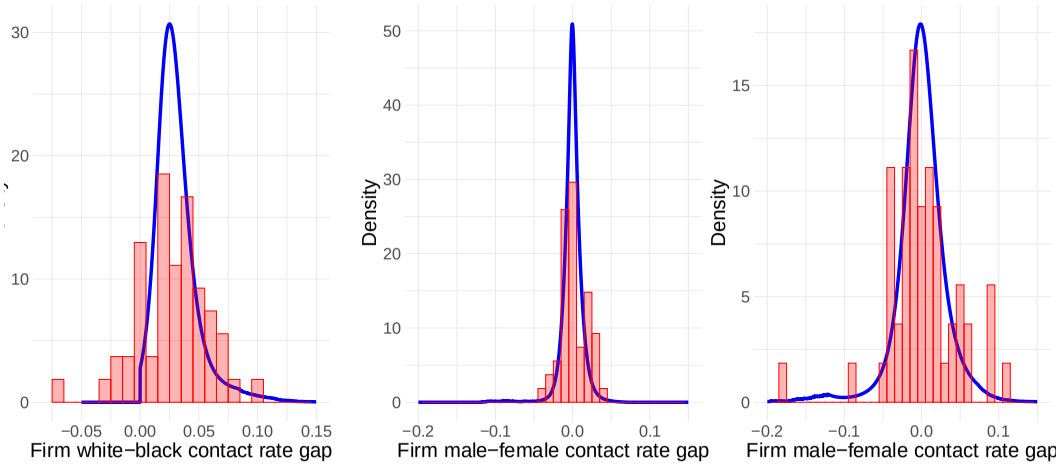
Figure E3: Variance-stabilized deconvolutions of racial discrimination without support restriction



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 an estimate of the population contact gap distribution constructed as in Panel (a) of Figure E1, but without the restriction that the density of Δ_f is weakly positive.

gure E4: Local deconvolutions of firm-level discrimination distributions

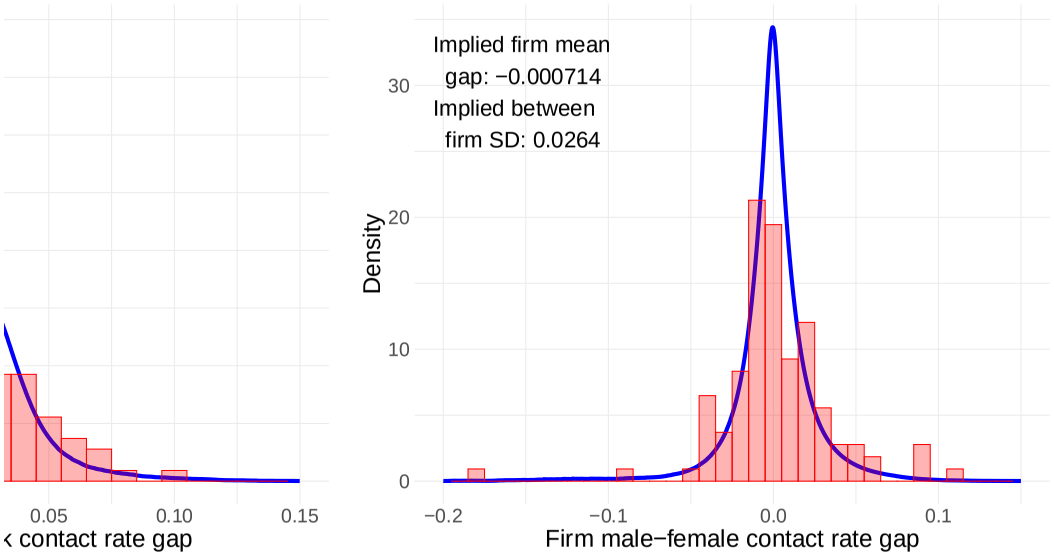
ce b) Gender



umetric estimates of the distribution of firm-specific contact gaps estimated separately for firms with above / (a) presents estimates for white-Black contact rate differences, and panel (b) presents estimates for male-female e distribution of estimated firm contact gaps in each group. Blue lines show estimates of population contact e population distributions are estimated by applying the deconvolveR package (Narasimhan and Efron, 2020) hin group, then numerically integrating over the group's empirical distribution of standard errors. A common e deconvolution step for both groups and calibrated so that the resulting marginal distribution matches the e estimate from Table 4. In panel (a), the density of population z-scores is constrained to be weakly positive

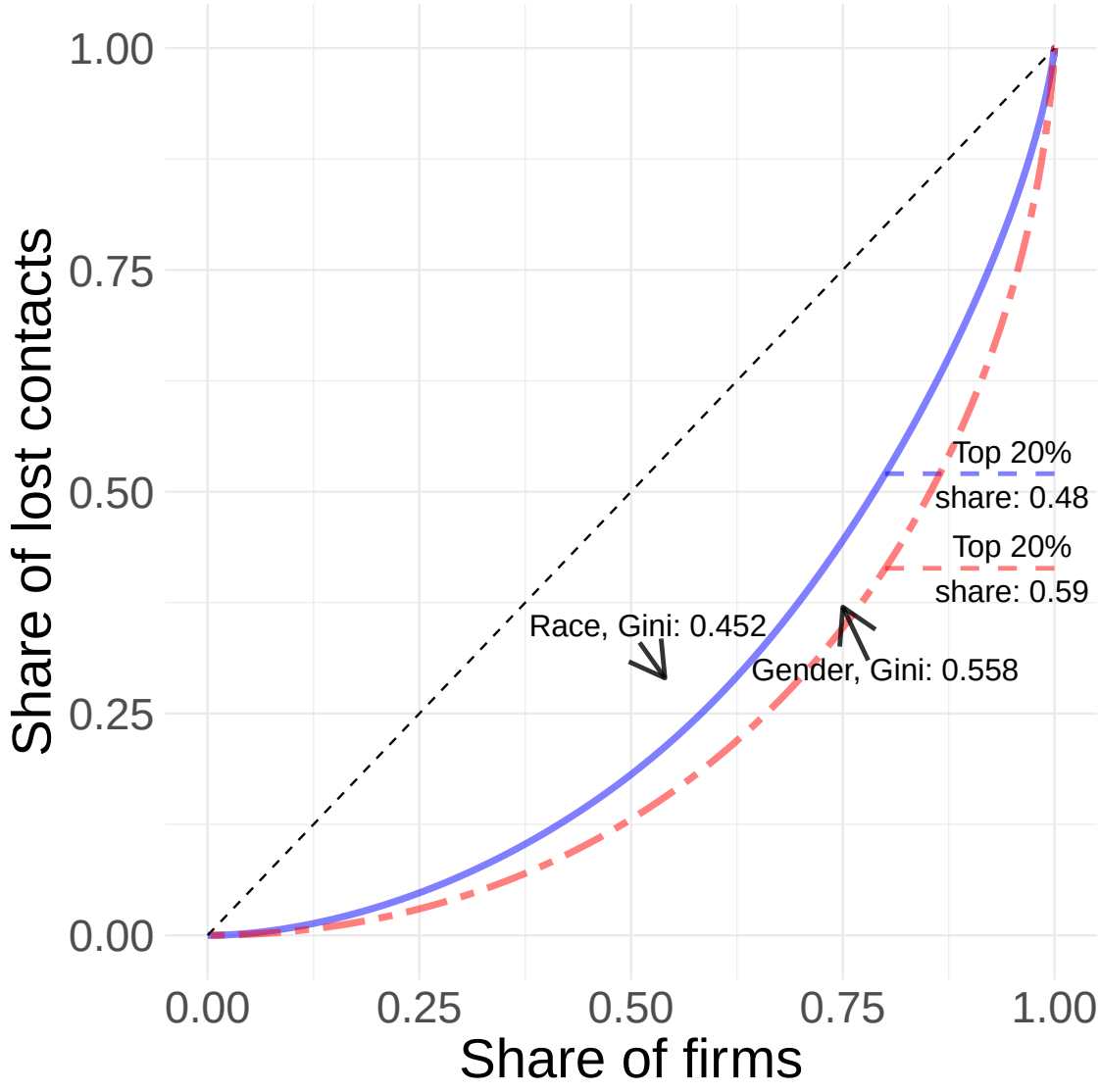
E5: Marginal distributions of firm-level discrimination from local approach

ce b) Gender



metric estimates of the marginal distribution of firm-specific contact gaps corresponding to the group-specific presents estimates for white-Black contact rate differences, and panel (b) presents estimates for male-female distribution of estimated firm contact gaps. Blue lines shows estimates of population contact gap distributions. the average of the group-specific densities in Figure E4.

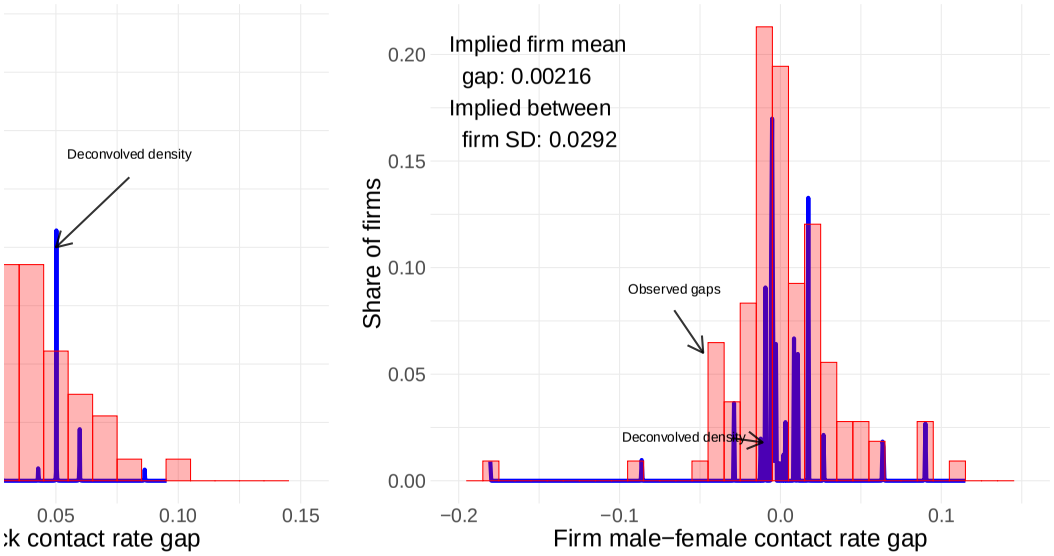
Figure E6: Local deconvolution Lorenz curves



Notes: This figure displays Lorenz curves implied by the non-parametric deconvolution estimates of race and gender contact gap distributions in Figure E5. 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.

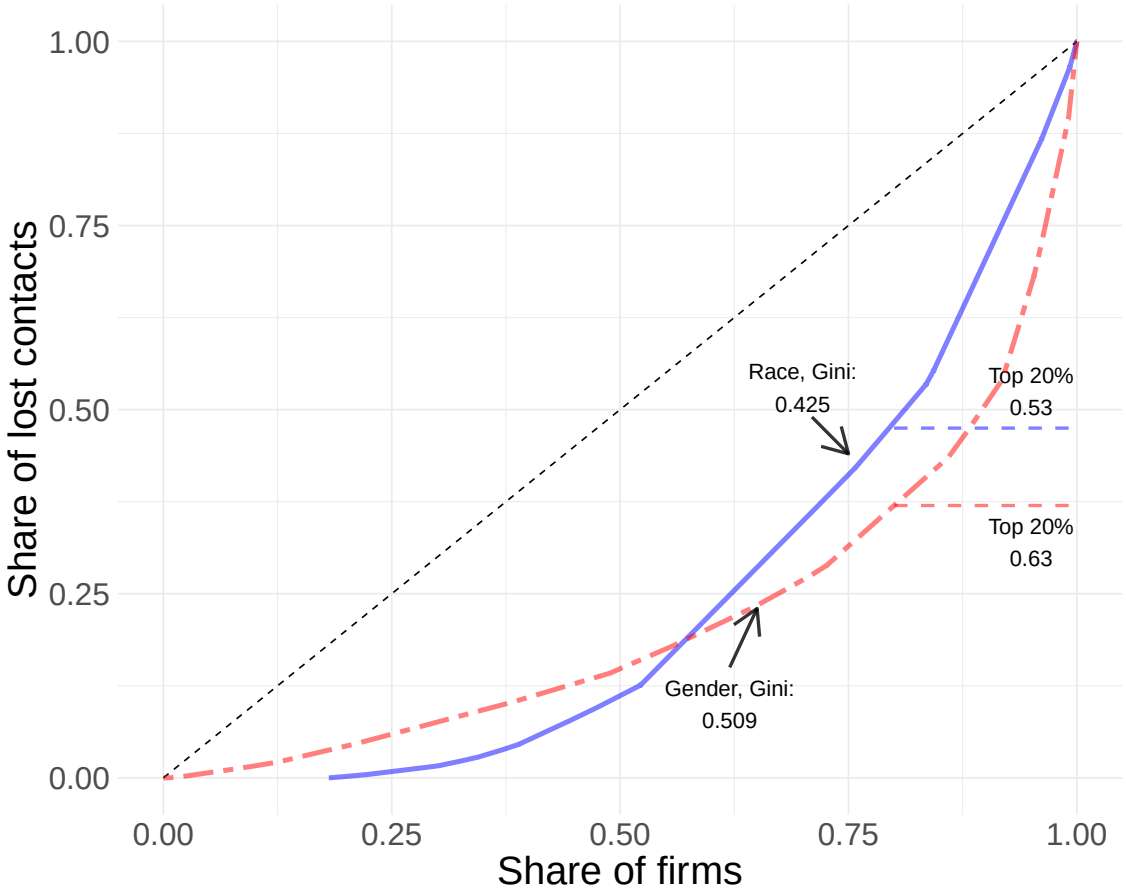
E7: NPMLE estimates of marginal distributions of firm-level discrimination

ce b) Gender



ametric maximum likelihood estimates of the distribution of firm-specific contact gaps estimated using the Panel (a) presents estimates for white-Black contact rate differences, where we impose the restriction that all l panel (b) presents estimates for male-female differences. Red histograms show the distribution of estimated estimates of population contact gap distributions. Population distributions are estimated allowing a non- the mixing distribution of contact gaps and standard errors. The figures plot the marginal distribution of is discrete, the blue lines plot the probability mass function in below, while the histogram reports the share of

Figure E8: NPMLE Lorenz curves



Notes: This figure displays Lorenz curves implied by the NPMLE estimates of race and gender contact gap distributions reported in Figure E7. 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. Linear interpolation has been used between mass points, which generates kinks in the curve. 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.

E-lit-001: How Black Are Lakisha and Jamal? Racial Perceptions from Names Used in Correspondence Audit Studies

S. Michael Gaddis | Sociological Science 4: 469-489 | 2017-09-06

https://sociologicalscience.com/download/vol-4/september/SocSci_v4_469to489.pdf

About this source

The paper shows that names are a noisy and non-neutral instrument for race, and that the noise is not random. Black first names commonly given by less-educated mothers (DaShawn, DaQuan, Tanisha, Lakisha) are perceived as black by more than 95 percent of respondents, while black names given by more-educated mothers (Nia, Jalen, Malia, Malcolm) fall far below that; in the raw data the gap is 91.8 percent versus 59.5 percent across the education quartiles, and in a logistic model holding the name's population blackness constant the lowest-education quartile carries an odds ratio of 2.84 and the highest 0.43. The same gradient runs in the opposite direction for white names (0.50 for low-education white names, 1.10 for high-education), so a low-status white name like Cheyanne reads white only 55.6 percent of the time. This cuts two ways for the section's hypothesis. It supports the class-confound critique — the strength of the racial signal and the lowness of the implied class position move together, so a study choosing the most 'reliably black' names is by construction choosing the most status-loaded ones, and the estimated gap mixes race with perceived class. But it equally supports an attenuation reading: 20 of the 80 black names were perceived as black by fewer than 75 percent of respondents even with a black surname attached, so audits built on weak names transmit the treatment only part of the time and understate the true race effect. Gaddis also documents that last names matter enormously and independently — a black first name with a white last name reads black only 66.5 percent of the time versus 82.5 percent with a black last name — and notes that a trio of otherwise similar audits of for-profit credentials reached conflicting conclusions, with the one that signaled race through white first names plus black surnames (Darolia et al. 2016) finding no discrimination at all. Finally, he shows how casually the field has treated this: 37.5 percent of correspondence audits simply reuse Bertrand and Mullainathan's names, and fewer than one in five pretested their names at all. Bertrand and Mullainathan's own small Chicago pretest, he notes, had already flagged that Maurice and Jerome did not read as strongly black.

Bound as the original PDF. Retrieved 2026-09-10. sha256 ed91904f1feb0556...

Bound into an Economic Briefing · by Claude, may contain mistakes

Full document begins on the next page.

E-lit-001: https://sociologicalscience.com/download/vol-4/september/SocSci_v4_469to489.pdf

sociological science

How Black Are Lakisha and Jamal? Racial Perceptions from Names Used in Correspondence Audit Studies

S. Michael Gaddis

University of California, Los Angeles

Abstract: Online correspondence audit studies have emerged as the primary method to examine racial discrimination. Although audits use distinctive names to signal race, few studies scientifically examine data regarding the perception of race from names. Different names treated as black or white may be perceived in heterogeneous ways. I conduct a survey experiment that asks respondents to identify the race they associate with a series of names. I alter the first names given to each respondent and inclusion of last names. Names more commonly given by highly educated black mothers (e.g., Jalen and Nia) are less likely to be perceived as black than names given by less educated black mothers (e.g., DaShawn and Tanisha). The results suggest that a large body of social science evidence on racial discrimination operates under a misguided assumption that all black names are alike, and the findings from correspondence audits are likely sensitive to name selection.

Keywords: racial discrimination; inequality; names; audit studies; experiments

MODERN social science evidence of racial discrimination stems mostly from a type of field experiment known as an audit study, which matches candidates on all characteristics except race to examine racial differences in outcomes. Originally developed in the 1960s to capture more elusive forms of racial discrimination in the post–Civil Rights era, modern audit studies have shifted from the in-person to the correspondence method, which uses names to signal the race of hypothetical subjects. With few exceptions, correspondence audits regularly find lower phone and/or email response rates for applications with black names compared to white names in both employment and housing (Gaddis 2015; Hanson et al. 2016; Hogan and Berry 2011). Researchers have continued to find evidence of racial discrimination even as correspondence audits have expanded to include a broader domain of actors, such as politicians, prospective roommates, public officials, and health care professionals (Butler and Broockman 2011; Einstein and Glick 2017; Gaddis and Ghoshal 2017; Giulietti, Tonin, and Vlassopoulos 2015; Sharma, Mitra, and Stano 2015).

Such correspondence audits enable researchers to circumvent a number of critiques of the in-person method (Heckman 1998; Heckman and Siegelman 1993), collect larger samples, and reduce research costs. However, researchers also lose the ability to more directly convey race through appearance and interaction and instead rely solely on names to signal race. The vast majority of the recent evidence on racial discrimination hinges on individuals' racial perceptions from these names.

An exhaustive search of correspondence audits (conducted by both mail and internet) that use names to signal race yields 89 studies in published or working

Citation: Gaddis, S. Michael. 2017. "How Black Are Lakisha and Jamal? Racial Perceptions from Names Used in Correspondence Audit Studies." Sociological Science 4: 469-489.

Received: May 18, 2017

Accepted: June 12, 2017

Published: September 6, 2017

Editor(s): Jesper Sørensen, Olav Sorenson

DOI: 10.15195/v4.a19

Copyright: © 2017 The Author(s). This open-access article has been published under a Creative Commons Attribution License, which allows unrestricted use, distribution and reproduction, in any form, as long as the original author and source have been credited.

paper form since 1970. The occurrence of these studies has accelerated in recent years (72 studies, or 81 percent, have been published or otherwise made available since 2010).[1] Researchers across a variety of disciplines—sociology, economics, political science, psychology, and management—have published these studies in some of the top generalist journals. To fully capture the gravity of how names inform a scientific understanding of racial prejudice, discrimination, and stereotypes, we can expand the search of the scientific literature beyond correspondence audits to also include laboratory, vignette, and other experiments, which easily surpasses 250 studies since the year 2000.[2]

Unfortunately, no research has systematically investigated the validity of using names to signal race. In correspondence audits, researchers seem to assume a consensus on what constitutes distinctively black and white names and that any one "black" name should yield similar results as any other "black" name. However, scientific explanations of how researchers select black and white names are woefully lacking. These explanations range from no details (10 percent) to selecting from published popular name lists (often with no racialized data) (17.5 percent), recycling Bertrand and Mullainathan's (2004) selected names (or those from Levitt and Dubner's 2005 list) (37.5 percent), using state birth record data without a pretest (17.5 percent), and using pretests to examine perceptions of names (17.5 percent). In other words, less than 1 out of every 5 studies using names to signal race has scientifically examined relevant data to see how people perceive race from the selected names. When researchers do gauge racial perceptions from names using pretests, they often have very small sample sizes and/or query only college students, and no researchers have used experimental manipulations. This is particularly troubling because name-based correspondence audits have become the leading method of providing evidence of racial discrimination.

Although the research base clearly shows that race can be signaled through names and that using names as a signal of race can successfully capture some version of racial discrimination, it is unclear whether people actually *perceive* the signal of race in the same way across the variety of names used in past research. At least 3 characteristics of a name may influence an individual's perception of race from that name: (1) the population racial composition of a first name, (2) the population socioeconomic status (SES) composition of a first name, and (3) the population racial composition of a last name. Although some racial discrimination research has been concerned with the second characteristic (Bertrand and Mullainathan 2004; Gaddis 2015), nearly none has paid attention to the first and third characteristics.

Studies that use names to signal race rely on the assumption that there are first names among the population that are unique to certain races. Thus, if a name, such as Jamal, more frequently belongs to a black person, the general population should recognize that name as black. Since at least the 1960s, a small percentage of black parents have chosen certain first names for their children at much higher frequencies than white parents, making these names racially distinctive (Fryer and Levitt 2004; Lieberson 2000; Lieberson and Mikelson 1995). However, the majority of black parents do not name their children using distinctive first names. For example, from 1994 to 2012 in New York state, only 15 names were used by black mothers more than 3,000 times. Most of these 15 names were commonly given by both black and

white mothers: Anthony, Ashley, Joshua, Kayla. Only one of these 15 names was more commonly given by black rather than white mothers: Isaiah. Overall, only 18.9 percent of black children born in New York during this period have a name that is racially distinctive as black (more commonly given by black rather than white mothers).

Moreover, data from New York suggest that the most distinctively black names are not exclusively used by blacks. For example, names such as Jamal and Latoya belong to children with black mothers at rates of 80 percent and 84 percent, respectively, leaving 20 percent of Jamals and 16 percent of Latoyas as nonblack or multiracial. Other "black" names such as Keisha and Leroy have lower rates (56 percent and 61 percent, respectively). All four of these names have been used to signal black applicants in multiple correspondence audits.

Taken together, these two issues should make it clear that, at best, first names can only be imperfect proxies of race. Researchers take a shortcut by first using a specific subset of names and then taking a continuous variable of racial naming practices and turning it into a binary (i.e., white name or black name). Thus, even if data on actual population naming practices by race could perfectly predict perceptions of race from these names, we would expect that, for instance, 20 percent of the time Jamal would be perceived as nonblack. However, individual perceptions may not perfectly align with reality, as one often-overlooked small sample pretest finding from Bertrand and Mullainathan (2004) suggests.[3]

A second important factor determining what racial cues a first name may signal is the correlation between parental SES and names. Using birth record data from California, Fryer and Levitt (2004) find that "[b]lacker names are associated with lower-income zip codes [and] lower levels of parental education" (p. 786). Moreover, although there are fewer instances of unique naming patterns among white parents, these unique names are still correlated with SES in the New York birth record data. Because both race and SES influence parental naming practices, the racial perception from a name may be biased by the SES-based naming practices. Whereas some find that individuals' racial perceptions from names are not biased by these SES-based practices (Bertrand and Mullainathan 2004), other research suggests this may not be true in all instances (Figlio 2005; Gaddis 2015; Kirschenman and Neckerman 1991).

Finally, researchers have focused minimal attention on last name selection with regard to race. However, publicly available U.S. Census data show that only nine last names among the most frequently occurring 1,000 are majority black, with another 15 last names registering at 40 percent to 50 percent black (U.S. Census Bureau 2012). Earlier correspondence audits often mixed "black" first names with both predominantly black last names and predominantly white last names within a study (Hanson and Hawley 2011; Milkman, Akinola, and Chugh 2012), whereas more recent studies often match first and last names by race (Gaddis 2016; Hanson et al. 2016). The effect of this selection on outcomes in correspondence audits is not only unknown but heretofore unquestioned.

One additional potential wrinkle in signaling race through names remains: the characteristics of the individual *receiving* the signal. For instance, we might expect that blacks would be more familiar with black names generally and thus more likely to recognize a particular name as black, independent of the SES associations

of that name. Other characteristics, such as an individual's age, gender, and SES might matter as well. Although no correspondence audits directly acknowledge this issue, the implications are clear. If whites are overrepresented in a particular audit context (e.g., real estate agents) and also less likely to receive the racial signals sent by researchers, we may underestimate true discrimination rates by using poorly performing names.

Overall, a body of research suggests that further examination of racial perceptions from names will substantially improve our understanding of existing discrepancies in racial discrimination research and help lead to future laboratory, field, and survey experiments with higher internal validity. Unfortunately, few studies have examined different perceptions of race from names and none do so in an experimental framework. The present research stands to make an important contribution to the social sciences by providing the first scientific evidence on perceptions of race from names. I proceed by conducting a survey experiment and examining a number of research questions on individual racial perceptions from names.

Research Questions

In this study, I address four primary research questions:

(1) Are individual perceptions of race from first names congruent with population-level naming practices and prior correspondence audits?
(2) Does the inclusion of different types of last names affect individual perceptions of race from first names?
(3) Do variations in population-level naming practices by race, education, and popularity affect individual perceptions of race from first names?
(4) Do respondent characteristics—particularly race, gender, age, and SES—affect individual perceptions of race from first names?

Data and Methods

Using Amazon's Mechanical Turk

To examine these research questions, I conducted a survey experiment using Amazon's Mechanical Turk (MTurk) from September 2014 through August 2015. MTurk is a crowdsourcing microtask marketplace where individuals can assign (requesters) or perform (workers) tasks (Human Intelligence Tasks, or HITs) for monetary compensation. MTurk has become popular among social scientists, particularly for conducting survey experiments in the fields of psychology, political science, sociology, and health, among others (Campbell and Gaddis forthcoming; Dowling and Miller 2016; Horne et al. 2015). Researchers have praised MTurk for its relatively low cost and quick turnaround for data and have offered cautious optimism regarding generalizability (Horton, Rand, and Zeckhauser 2011; Weinberg, Freese, and McElhattan 2014). Moreover, on a number of dimensions, MTurk represents a superior alternative to using undergraduate students, a ubiquitous sample in experimental psychology (Sears 1986).

Internal and External Validity with MTurk

Scholars have raised two primary issues of concern about MTurk: (1) respondent demographics and representativeness, or external validity; and (2) the reliability of reporting and data quality, or internal validity. Numerous recent studies address these issues in depth. First, scholars have presented significant evidence that shows that although MTurk's pool of participants are not demographically the same as the U.S. population as a whole, the participants are also not as homogeneous as one might imagine nor radically different from the U.S. population (Berinsky, Huber, and Lenz 2012). Compared to the U.S. population, MTurk samples are composed of slightly more women and are younger and more educated but with similar income distributions to the U.S. population (Buhrmester, Kwang, and Gosling 2011; Ipeirotis 2010; Paolacci, Chandler, and Ipeirotis 2010). MTurk may also slightly underrepresent blacks and Hispanics while overrepresenting Asians (Berinsky et al. 2012; Chandler, Mueller, and Paolacci 2014). Workers on MTurk also lean towards more liberal attitudes and opinions (Berinsky et al. 2012). There is some evidence that these demographic differences account for minimal differences in effect sizes between MTurk and other Internet survey platforms that claim representative samples (Weinberg et al. 2014). Moreover, careful checks of moderating demographic variables that are not representative of the United States in MTurk samples and/or weighting may alleviate concerns regarding external validity (Mullinix et al. 2016; Weinberg et al. 2014).

Second, concerns regarding internal validity may be exaggerated. Research suggests that data from MTurk are more reliable than undergraduate lab samples and, at minimum, equal to other Internet samples (Behrend et al. 2011; Buhrmester et al. 2011; Paolacci et al. 2010; Weinberg et al. 2014). A number of studies use "catch trials" and/or longitudinal samples to verify respondent demographics and attention to the tasks at hand (Mason and Suri 2012; Rand 2012). Additionally, Peer, Vosgerau, and Acquisti (2014) find that limiting HITs to high-reputation workers (those with $\geq$95 percent HIT approval ratings) is sufficient to maintain high data quality, and adding catch trials or attention-check questions do not improve data quality further.

Sample and Survey Setup

Requesters on MTurk can list requirements of workers and block anyone not meeting those requirements from accessing the HIT. I created a sampling frame within MTurk by limiting workers to only those with a U.S. address and, following the findings of Berinsky et al. (2012), further restricted the sample to only those workers with a HIT approval rate $\geq$95 percent to improve data quality.

I published a batch of 150 to 200 HITs approximately every week during the 12-month period. After each assignment closed, I flagged previous participants so they could neither repeat nor see new batches of the survey. Additionally, I frequently checked the major MTurk message boards to be sure that no major discussions were taking place online that disclosed the nature of the experimental manipulation. I chose to open small batches of the survey experiment at multiple intervals rather than post a single large batch with thousands of assignments so that

the experiment was always relatively fresh to anyone who had not yet completed the survey. Respondents who accepted the HIT were taken to a redirect page written in JavaScript that randomized the survey condition by randomly sending the respondent to one of 50 different survey webpages to complete the survey. The first treatment assignment phase directed respondents to one of 10 sets of first names (each set contained 20 different first names), with the goal of 10 percent to each. The second treatment assignment phase directed respondents to either first names only or one of four last name sets, with the goal of 30 percent to first names only and 17.5 percent to each of the last name conditions (within each assignment from phase 1). In total, 8,424 respondents started the survey and 7,881 completed the survey (93.5 percent).

Selecting Names to Test

I selected names for this study using New York state birth record data for all births from 1994 to 2012 obtained from the New York State Department of Health to examine population-level race and SES characteristics.[4] These data separately list the total number of births by (1) name and mother's race and (2) name and mother's education. This data structure allowed me, for example, to choose two names similar in terms of mother's race but different in terms of mother's education—in other words, a black lower-SES name and a black middle- to upper-SES name. Two examples used in this study are DaQuan and Jabari; 91.8 percent of children named DaQuan and 92.1 percent of children named Jabari are born to black mothers. These names are equal in blackness but vary by mother's education; only 12.8 percent of mothers who name their child DaQuan have some college or more education, whereas 56.8 percent of mothers who name their child Jabari have some college or more education.

Additionally, when possible, I selected names that were used in previous or ongoing audit studies from different disciplines (e.g., Bertrand and Mullainathan 2004; Gaddis 2015; Milkman et al. 2012). This will permit me, in future work, to compare how different racial response rates by individual names in those studies match with the racial perception rates of this study. In total, I use 200 first names; 10 different sets of 20 names, 80 of which are black (>50 percent born to black mothers in the New York data), 80 of which are white (>50 percent born to white mothers in the New York data), and 40 of which are Hispanic.[5] Within each set of 20 first names, 8 are black, 8 are white, and 4 are Hispanic. For brevity and because fewer correspondence audits to date have examined outcomes using Hispanic names, I do not include any additional results for Hispanic first names in this article (see Gaddis 2017). Table 1 shows the full list of names in each set, and Table A1 in the online supplement shows the full list of names by mother's race, mother's education, and total frequency in the New York state birth record data.

Finally, I chose last names using frequently occurring surnames from the 2000 Census, which lists the population racial composition of last names in the United States (U.S. Census Bureau 2012). I show the population-level racial occurrence of these last names in Table 2.

Table 1: Sets of names used in the survey experiment.

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9	Set 10
1	*Jalen*	DeAndre	*Malia*	Tanisha	*Jabari*	Taniya	*Darius*	*Janae*	Hakim	Kenya
2	Lamar	*Reginald*	Monique	**Latoya**	Keyshawn	Divine	Maurice	Heaven	Tremayne	**Lakisha**
3	**DaQuan**	Marques	Shania	Aisha	Dwayne	Keyana	Jayvon	*Erykah*	Delroy	Precious
4	*Nia*	*Cedric*	*Jada*	Rasheed	Marquise	*Bria*	Latrell	Octavia	Jerome	Kiara
5	Ebony	Aaliyah	**DeShawn**	Darnell	Tamia	*Savion*	Tionna	Denzel	Keisha	Kareem
6	Shanice	Denisha	*Malcom*	Tyrone	Kimani	D'Andre	Lyric	Terrell	**Latonya**	Leroy
7	Tyra	Jasmine	*Quincy*	Jamal	Aniya	Jaleel	*Kaylah*	Wendell	**Tamika**	Terell
8	**Unique**	Chanel	*Andre*	Jermaine	*Amari*	Kevon	Sade	Tevin	Ashanti	Deshawn
9	*Caleb*	Brian	*Mary*	Joan	*Molly*	Katie	**Mayer**	Cheyanne	*Brett*	*Anne*
10	Charlie	Cody	Lisa	**Melany**	Amy	*Madeline*	**Edwin**	*Chloe*	*Matthew*	Carrie
11	**Ronny**	*Ethan*	Barbara	Hilary	*Claire*	*Emily*	*Graham*	*Margaret*	Steven	Kristen
12	*Aubrey*	*Zachary*	Stephanie	Todd	*Katelyn*	*Abigail*	*August*	*Harper*	Robert	*Meredith*
13	Erica	*Megan*	*Ryan*	*Geoffrey*	*Jake*	*Wyatt*	**Angie**	**Irvin**	*Allison*	*Brendan*
14	**Lesly**	Susan	Seth	Jay	*Logan*	Dustin	Marlene	**Jordy**	*Jill*	*Neil*
15	**Brenda**	Deborah	*Maxwell*	Brad	*Connor*	Luke	Cassie	**Edgar**	Laurie	Daniel
16	Heidi	*Erin*	*Spencer*	Greg	Scott	Hunter	*Charlotte*	*Finn*	*Sarah*	*Paul*

Note: Names 1 through 8 correspond to population-level names that are mostly black; names 9 through 16 correspond to population-level names that are mostly white. **Bold** indicates the name is in the lowest quartile of mother's education within race and *italics* indicate the name is in the highest quartile of mother's education within race. Kiara and Jasmine are not majority nor plurality black in the New York birth record data. Both Jasmine and Kiara are listed on Levitt and Dubner's (2005) top black female names list, increasing the likelihood that either name could be used as a black name in audit studies of racial discrimination. Additionally, Jasmine has been used as a black name in a previous audit study (Jacquemet and Yannelis 2012). Thus, I include both as black names in this study.

Table 2: Last names used in the survey experiment.

Name	Rank	Percent White	Percent Black	Percent Hispanic
Washington	138	5.2%	89.9%	1.5%
Jefferson	594	18.7%	75.2%	1.6%
Booker	902	30.0%	65.6%	1.5%
Banks	278	41.3%	54.2%	1.5%
Jackson	18	41.9%	53.0%	1.5%
Mosley	699	42.7%	52.8%	1.5%
Becker	315	96.4%	0.5%	1.4%
Meyer	163	96.1%	0.5%	1.6%
Walsh	265	95.9%	1.0%	1.4%
Larsen	572	95.6%	0.4%	1.5%
Nielsen	765	95.6%	0.3%	1.7%
McGrath	943	95.9%	0.6%	1.6%
Stein	720	95.6%	0.9%	1.6%
Decker	555	95.4%	0.8%	1.7%
Andersen	954	95.5%	0.6%	1.7%
Hartman	470	95.4%	1.5%	1.2%
Orozco	690	3.9%	0.1%	95.1%
Velazquez	789	4.0%	0.5%	94.9%
Gonzalez	23	4.8%	0.4%	94.0%
Hernandez	15	4.6%	0.4%	93.8%

Note: Data from U.S. Census Bureau (2012).

Survey Questions

The pertinent survey question on racial perception asked: "For each of the following names, list the race or ethnicity that you associate with that name (for example: white, black, Hispanic, Latino, Asian, etc.).[6] If you do not have a clear racial or ethnic association with a name, you may type 'none'." Open-ended responses were recoded to indicate Asian (1.7 percent of all responses), black (40.7 percent), Hispanic or Latino (8.4 percent), white (44.2 percent), none (4.8 percent), or other (0.2 percent). When multiple racial and/or ethnic categories were suggested by the respondent, I used the first word typed as the primary perception; only 0.3 percent of all answers indicated some form of bi- or multiracial perception.[7]

In addition to the questions on racial perceptions from names, the first page of the survey included nine demographic and background questions about the survey respondent: age, race and/or ethnicity, sex, relationship status, any children less than 18, highest level of education completed, combined household income, employment status, and zip code. The second page included the name questions, and the final page of the survey asked for the respondent's MTurk ID for verification and payment.

Methods of Analysis

In the first part of the next section, I present basic bivariate descriptive results that show the perception of names as either "congruent" or "incongruent" with population-level naming practices and prior correspondence audits. This approximates the real-world process that occurs during field experiments. For example, a respondent's recoded response of perception as "black" is congruent with a name used to signal a black applicant in previous correspondence audits, but a recoded response of perception as "none," "Asian," or "white" is incongruent with the same name. The match between a hypothetical researcher's *intended signal* of race and a respondent's *perceived* racial identification of that person is what matters in a field experiment. For simplification, I refer to a matched perception with the intended signal as "congruent perception" at the individual level and "congruent perception rate" at the aggregate level in the remainder of this article.

In the second part of the next section, I use logistic regression to model the effects of name characteristics and respondent characteristics in predicting odds ratios of congruent perceptions of names:

$$ln(p/[1-p]) = \alpha_n + \beta_1 X_n + \beta_2 V_r \quad (1)$$

In the equation above, α_n is the name-level intercept, X_n represents a vector of name variables (gender, last name, mother's education quartile, mother's race percentage, and total frequency), and V_r represents a vector of respondent variables (age, race and/or ethnicity, sex, relationship status, any children less than 18, highest level of education completed, combined household income, and employment status). Mother's race percentage (black or white) is a continuous variable, whereas mother's education is transformed into race-specific quartiles of the percentage of mothers with some college or more education. These models include cluster-corrected standard errors at the respondent level.

Results

Descriptive Results

The first two figures show the congruent perception rates of all individual black (Figure 1) and white (Figure 2) names by type of last name (white last name included, black last name included, and no last name). Each bar indicates the percentage of survey respondents whose perception of a name matches the signal. Names are sorted in ascending order by racially matched last name perception rates. The primary discussion in this section focuses on congruent perception rates when names are matched with the appropriate race last name. Generally, the congruent perception rates are lower when first names are given no last name or a racially mismatched last name. Detailed discussion of these aspects follows in the regression results section.

There is much variation in congruent perception rates within each set of racialized names. For black names (Figure 1), the respondents were least likely to congruently perceive the names Bria, Sade, Kaylah, Lyric, and Jasmine[8] when matched

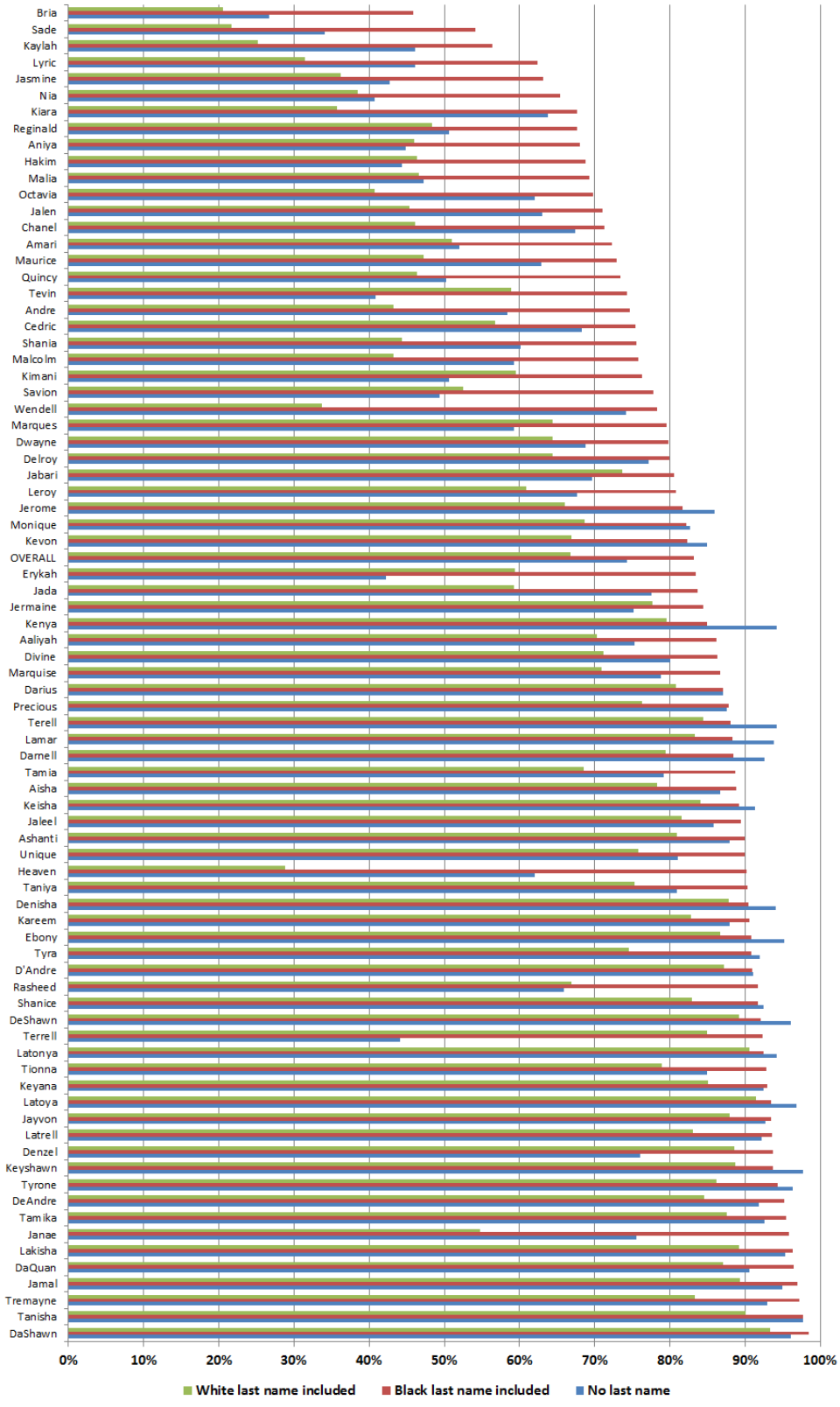


Figure 1: Congruent perception rates of black first names by last name status.

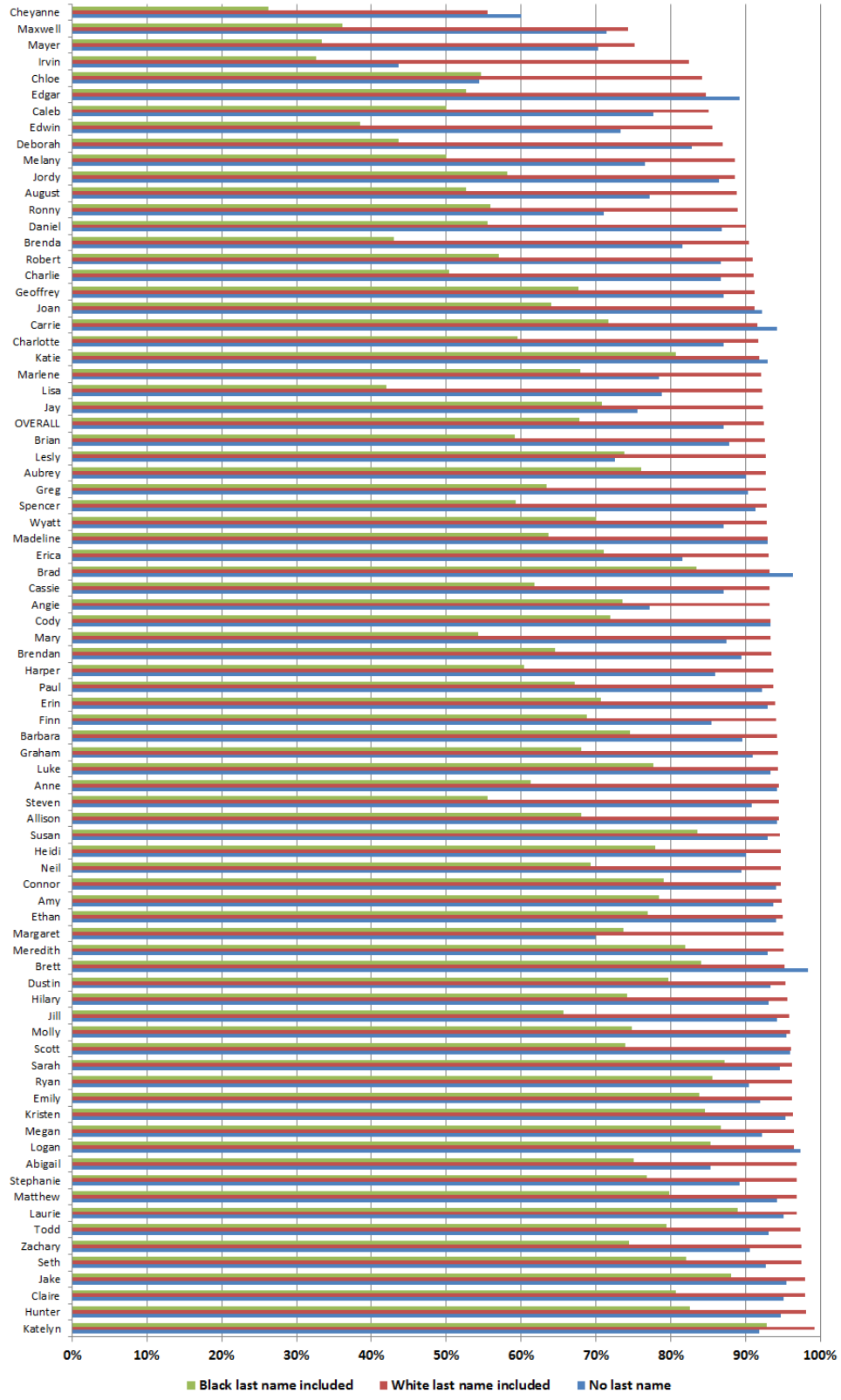


Figure 2: Congruent perception rates of white first names by last name status.

with a black last name. All of these names were perceived as black among less than 65 percent of the respondents. Even when black last names were included, 20 of the 80 black names (25 percent) were perceived as black among less than 75 percent of the respondents, indicating poor choices for use in experiments signaling race. Conversely, congruent perception rates were quite high for the names DaShawn, Tanisha, Tremayne, Jamal, and Daquan. All of these names were perceived as black among more than 95 percent of the respondents. When black last names were included, 30 of the 80 black names (37.5 percent) were perceived as black among more than 90 percent of the respondents, indicating very good choices for use in field experiments signaling race. The congruent perception rate across all black names is 75.0 percent when given no last name, 82.5 percent when given a black last name, and 66.5 percent when given a white last name.

For white names (Figure 2), the respondents were least likely to congruently perceive the names Cheyanne, Maxwell, Mayer, Irvin, and Chloe when matched with a white last name. All of these names were perceived as white among less than 85 percent of the respondents. Only Cheyanne (55.6 percent) and Maxwell (74.3 percent) were perceived as white among less than 75 percent of the respondents, indicating poor choices for use in field experiments signaling race. Conversely, congruent perception rates were extremely high for the names Katelyn, Hunter, Claire, Jake, and Seth. All of these names were perceived as white among more than 97 percent of the respondents. When white last names were included, 67 of the 80 white names (83.8 percent) were perceived as white among more than 90 percent of the respondents, indicating very good choices for use in experiments signaling race. The respondent congruent perception rate for all white names is 87.3 percent when given no last name, 92.4 percent when given a white last name, 67.9 percent when given a black last name, and 17.8 percent when given a Hispanic last name.

One other significant variation that stands out from these figures comes from differences by mothers' education. Respondents are much more likely to congruently perceive a black name from mothers with *lower* education levels like DaShawn, DaQuan, or Lakisha rather than from mothers with *higher* education levels like Nia, Malcolm, or Malia. Respondents are also much more likely to congruently perceive a white name from mothers with *higher* education levels like Claire, Jake, or Abigail rather than from mothers with *lower* education like Cheyanne, Irvin, or Jordy, although the patterns for white names do not appear as strong as for black names.

Table 3 delves into the raw data and confirms the patterns discussed above. Generally, respondents congruently perceive white names at higher rates than others except when matched with a Hispanic last name. Respondents also congruently perceive male and female white names at equal rates but have less trouble congruently perceiving male rather than female black names. Finally, there's some evidence that black and white respondents can congruently perceive names that match their own race more readily that those of other races.

Table 3: Descriptive statistics: congruent perception.

	Signaled Race of Name	
	White	Black
Overall congruent perception	73.4%	74.7%
Characteristics of name		
With no last name	87.3%	75.0%
With racially matched last name	92.4%	82.5%
With white last name (black first only)	—	66.5%
With black last name (white first only)	67.9%	—
With Hispanic last name (white first only)	17.8%	—
Male	73.0%	77.3%
Female	73.7%	71.9%
Lowest quartile of mother's education	64.0%	91.8%
Middle two quartiles of mother's education	73.2%	76.8%
Highest quartile of mother's education	75.7%	59.5%
Characteristics of respondent		
White	74.7%	75.1%
Black	67.7%	79.6%
Hispanic	69.8%	75.4%
Asian	69.1%	67.7%
Male	73.1%	75.0%
Female	73.6%	74.3%
Total *N*	63,048	63,048
Respondent *N*	7,881	7,881

Note: Each respondent was asked to identify 16 different white and black names (see Table 1).

Regression Analyses

Table 4 shows logistic regression analyses predicting congruent perception of a black first name for 63,048 cases (7,881 respondents x 8 black names per respondent). The first model only controls for name characteristics and the second model also includes respondent characteristics. Model 1 shows that respondents are more likely to congruently perceive black names that are in the lowest quartile of mother's education (odds ratio [OR] = 2.84), when a black last name is included (OR = 1.63), and when a name is more black among the New York population of mothers (OR = 3.73). Respondents are less likely to congruently perceive black names when the name is female (OR = 0.72), when a white last name is included (OR = 0.65), when the name is in the highest quartile of mother's education (OR = 0.43), and when a name is more popular in the aggregate New York birth population (OR = 0.96). In the second model, we see that the respondent's own characteristics matter. None of the odds ratios for the name characteristics change substantially, but the results suggest that black respondents are more likely to congruently perceive black names

(OR = 1.43), as well as individuals in the middle age categories (age 25 to 34: OR = 1.30; age 35 to 49: OR = 1.32) and individuals with a household income of $25,000 to $49,999 (OR = 1.09) or $75,000 and higher (OR = 1.20).

Table 5 shows logistic regression analyses predicting congruent perception of a white first name for 63,048 cases. Both models control for various characteristics of the name itself, and the second model includes controls for the individual respondent's characteristics. Model 1 shows that respondents are more likely to congruently perceive white names when a name is female (OR = 1.08), when a white last name is included (OR = 1.80), when a name is in the highest quartile of mother's education (OR = 1.10), when a name is more white among the New York population of mothers (OR = 2.18), and when a name is more popular in the aggregate New York birth population (OR = 1.04). Respondents are less likely to congruently perceive white names when a black last name is included (OR = 0.30), when a Hispanic last name is included (OR = 0.03), and when a name is in the lowest quartile of mother's education (OR = 0.50). In the second model, we see that the respondent's own characteristics matter somewhat. None of the odds ratios for the name characteristics change substantially, but the results suggest that blacks (OR = 0.62), Hispanics (OR = 0.77), and individuals with a bachelor's degree or greater (OR = 0.83) are less likely to congruently perceive a white name.

Robustness Analyses

One alternative strategy is to model the dependent variable as whether or not the respondent simply perceives a name as black or white instead of whether the respondent congruently perceives a name as black or white. This strategy does not completely mimic the correspondence audit process of sending *and* receiving a signal because the choice of what signal sent is omitted. In these models I run regressions similar to model 2 presented in Table 4 and Table 5 separately for the sets of 80 black names and 80 white names. The results are very similar to those presented above (see Table A2 in the online supplement). Of note, the models suggest that the "more black" a white name is (based on the percentage of black mothers in the New York data naming their child that name), the more likely a respondent is to perceive that name as black; conversely, the "more white" a black name is (based on the percentage of white mothers in the New York data naming their child that name), the more likely a respondent is to perceive that name as white.

Discussion and Conclusion

In the 1964 Supreme Court case *Jacobellis v. Ohio*, Justice Potter Stewart famously said "I know it when I see it" in reference to what constitutes pornography. Many scholars seemingly have taken the same tactic when deciding what constitutes a black name: they know it when they see (or hear) it. The underlying assumption that black names comprise a uniform body that signals the same information has dominated the leading method used to investigate racial discrimination since the early 2000s. However, the present research shows that this assumption fails to hold

Table 4: Logistic regressions predicting congruent perception of a black first name.

	(1)	(2)
Name Characteristics		
Female name	0.724†	0.724†
	(0.015)	(0.015)
Last name (reference: no last name)		
Black last name included	1.632†	1.639†
	(0.054)	(0.054)
White last name included	0.653†	0.651†
	(0.020)	(0.020)
Quartile of mother's education (reference: middle two quartiles)		
Lowest quartile of mother's education	2.839†	2.852†
	(0.159)	(0.160)
Highest quartile of mother's education	0.433†	0.430†
	(0.010)	(0.010)
Percent black mothers in New York data	3.728†	3.915†
	(0.370)	(0.387)
Total number of births in New York data (ln)	0.960†	0.960†
	(0.012)	(0.012)
Respondent Characteristics		
Black (reference: white)		1.429†
		(0.080)
Hispanic		1.096
		(0.062)
Female		0.976
		(0.028)
Age: 25–34 (reference: 18–24)		1.295†
		(0.050)
Age: 35–49		1.316†
		(0.060)
Age: 50+		1.006
		(0.056)
Education: some college (reference: ≤HS)		1.058
		(0.051)
Education: Associate's degree or other 2-year degree		1.043
		(0.061)
Education: Bachelor's degree or higher		1.096
		(0.053)
Income: $25,000–$49,999 (reference: <$25,000)		1.089*
		(0.043)
Income: $50,000–$74,999		1.074
		(0.048)
Income: $75,000+		1.197†
		(0.054)
Constant	1.980†	1.363*
	(0.264)	(0.200)
N	63,048	63,048

Note: Odds ratios shown. Regressions also control for respondent's marital status, employment status, and whether respondent has any children under the age of 18. Cluster-corrected (respondent level) standard errors. †$p < 0.01$, *$p < 0.05$.

Table 5: Logistic regressions predicting congruent perception of a white first name.

	(1)	(2)
Name Characteristics		
Female name	1.079†	1.079†
	(0.025)	(0.025)
Last name (reference: no last name)		
White last name included	1.795†	1.812†
	(0.089)	(0.090)
Black last name included	0.302†	0.301†
	(0.013)	(0.013)
Hispanic last name included	0.030†	0.029†
	(0.001)	(0.001)
Quartile of mother's education (reference: middle two quartiles)		
Lowest quartile of mother's education	0.502†	0.502†
	(0.019)	(0.019)
Highest quartile of mother's education	1.098†	1.096†
	(0.030)	(0.030)
Percent white mothers in New York data	2.183†	2.238†
	(0.280)	(0.288)
Total number of births in New York data (ln)	1.044†	1.043†
	(0.011)	(0.011)
Respondent Characteristics		
Black (reference: white)		0.620†
		(0.035)
Hispanic		0.773†
		(0.043)
Female		1.020
		(0.035)
Age: 25–34 (reference: 18–24)		0.991
		(0.046)
Age: 35–49		1.021
		(0.057)
Age: 50+		1.121
		(0.080)
Education: some college (reference: ≤HS)		0.905
		(0.056)
Education: Associate's degree or other 2-year degree		0.935
		(0.068)
Education: Bachelor's degree or higher		0.832†
		(0.050)
Income: $25,000–$49,999 (reference: <$25,000)		0.959
		(0.047)
Income: $50,000–$74,999		0.903
		(0.049)
Income: $75,000+		0.967
		(0.053)
Constant	2.681†	3.076†
	(0.354)	(0.473)
N	63,048	63,048

Note: Odds ratios shown. Regressions also control for respondent's marital status, employment status, and whether respondent has any children under the age of 18. Cluster-corrected (respondent level) standard errors. $^{\dagger}p < 0.01$, $^{*}p < 0.05$

up when put under the scientific microscope. Indeed, black names used in previous correspondence audits vary significantly by individual perceptions of race. I find that a number of characteristics of an individual name matter: gender, popularity, type of last name included, and the average level of education of mothers who commonly give that name, among others.

The immediate implications of these findings are obvious: researchers can use this information to select names that signal race more clearly in correspondence audits. Whether researchers select the best-performing names among those I tested or conduct their own pretests before embarking on future correspondence audits, internal validity should increase in future racial discrimination research.

However, we should also question what these results might mean for the current body of discrimination research that is mostly not based on scientific selection of names to signal race. Differences in racial perceptions from names might explain differences in outcomes within and between correspondence audits. A recent trio of correspondence audits highlights this possibility (Darolia et al. 2016; Deming et al. 2016; Gaddis 2016). Three sets of researchers separately examined the effects of for-profit versus not-for-profit educational credentials in the labor market for black and white job candidates. Despite conducting the three correspondence audits during similar time periods with similar research questions and across many of the same cities, the findings regarding racial discrimination were quite different. Each chose different names to signal race, with one going the unique route of using generic or "white" first names coupled with "black" last names to signal a black applicant and found no evidence of racial discrimination (Darolia et al. 2016). Although other differences between these studies exist, the possibility that the racial signal from names might influence correspondence audit outcomes warrants further investigation.

Notes

1 This search, performed in September 2016, examined citation networks from high-profile audit studies, searches through the National Bureau of Economic Research and the Social Science Research Network, and personal correspondence between the author and a number of researchers conducting these studies.

2 A database of this research will be available in the near future at http://www.auditstudies.com.

3 Bertrand and Mullainathan (2004) found that some distinctively black names, Maurice and Jerome in particular, were not perceived as strongly black in a small test sample in Chicago.

4 The choice of New York birth record data is one of convenience. To my knowledge, no national-level data are available. The only other available large-scale, multiyear birth record data come from California. These data are expensive to obtain. Additionally, the racial demographics of New York are closer than California to the national percentages (e.g., blacks are 13.2 percent of the population nationally, 15.9 percent of the population in New York, and only 6.5 percent of the population in California). 1994 through 2012 was the full set of years available from New York at the beginning of this project. Although racial and SES-based naming practices may vary somewhat across regions, the question of importance is whether racial perceptions from names vary across regions. In supplemental analyses, I test whether respondents from New York vary from respondents in

the rest of the United States. I find no substantive differences in these analyses (available from author upon request), suggesting that the use of New York data likely has no significant bearing on the results.

5 I treat two names as black even though the New York data show that a plurality of mothers are white who name their children these two names: Jasmine and Kiara. In the case of Jasmine, one previous correspondence audit used the name to signal a black person (Jacquemet and Yannelis 2012). Additionally, both Jasmine and Kiara are listed on Levitt and Dubner's (2005) top black female names list, increasing the likelihood that either name could be used as a black name in audit studies of racial discrimination.

6 The survey question asks about "race or ethnicity." For brevity, I refer to this simply as race, even when discussing Hispanic ethnicity.

7 Dropping multiracial responses or treating them as "other" does not affect the substantive findings.

8 The low congruent perception rate of Jasmine is expected because only 33 percent of Jasmines in New York are born to black mothers. Again, because a previous correspondence audit used the name to signal a black person (Jacquemet and Yannelis 2012), I treat the name as black in this study.

References

Behrend, Tara S., David J. Sharek, Adam W. Meade, and Eric N. Wiebe. 2011. "The Viability of Crowdsourcing for Survey Research." *Behavior Research Methods* 43(3):800-13. https://doi.org/10.3758/s13428-011-0081-0

Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk." *Political Analysis* 20(3):351-68. https://doi.org/10.1093/pan/mpr057

Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *The American Economic Review* 94(4):991-1013. https://doi.org/10.1257/0002828042002561

Buhrmester, Michael, Tracy Kwang, and Samuel D. Gosling. 2011. "Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?" *Perspectives on Psychological Science* 6(1):3-5. https://doi.org/10.1177/1745691610393980

Butler, Daniel M., and David E. Broockman. 2011. "Do Politicians Racially Discriminate Against Constituents? A Field Experiment on State Legislators." *American Journal of Political Science* 55(3):463-77. https://doi.org/10.1111/j.1540-5907.2011.00515.x

Campbell, Colin S., and S. Michael Gaddis. Forthcoming. "'I Don't Agree with Giving Cash': A Survey Experiment Examining Support for Public Assistance." *Social Science Quarterly.* http://onlinelibrary.wiley.com/doi/10.1111/ssqu.12338/abstract

Chandler, Jesse, Pam Mueller, and Gabriele Paolacci. 2014. "Nonnaiveté among Amazon Mechanical Turk Workers: Consequences and Solutions for Behavioral Researchers." *Behavior Research Methods* 46(1):112-30. https://doi.org/10.3758/s13428-013-0365-7

Darolia, Rajeev, Cory Koedel, Paco Martorell, Katie Wilson, and Francisco Perez-Arce. 2016. "Race and Gender Effects on Employer Interest in Job Applicants: New Evidence from a Resume Field Experiment." *Applied Economics Letters* 23(12):853-6. https://doi.org/10.1080/13504851.2015.1114571

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. https://doi.org/10.1257/aer.20141757

Dowling, Conor, and Michael G. Miller. 2016. "Experimental Evidence on the Relationship Between Interest Group Funding and Candidate Vote Share." *Journal of Experimental Political Science* 3(2):152-63. https://doi.org/10.1017/XPS.2016.5

Einstein, Katherine Levine, and David M. Glick. 2017. "Does Race Affect Access to Government Services? An Experiment Exploring Street-Level Bureaucrats and Access to Public Housing." *American Journal of Political Science* 61(1):100-16. https://doi.org/10.1111/ajps.12252

Figlio, David N. 2005. "Names, Expectations and the Black-White Test Score Gap." NBER Working Paper No. 11195.

Fryer, Roland G., and Steven D. Levitt. 2004. "The Causes and Consequences of Distinctively Black Names." *Quarterly Journal of Economics* 119(3):767-805. https://doi.org/10.1162/0033553041502180

Gaddis, S. Michael. 2015. "Discrimination in the Credential Society: An Audit Study of Race and College Selectivity in the Labor Market." *Social Forces* 93(4):1451-79. https://doi.org/10.1093/sf/sou111

Gaddis, S. Michael. 2016. "A Field Experiment on Associate Degrees and Certificates: Statistical Discrimination, Stigma, Signal Boost, and Signal Saturation." Working paper.

Gaddis, S. Michael. 2017. "Racial/Ethnic Perceptions from Hispanic Names: Selecting Names to Test for Discrimination." Working paper. Retrieved August 8, 2017. https://ssrn.com/abstract=2975829

Gaddis, S. Michael, and Raj Ghoshal. 2017. "A Racial Stratification System for the 21st Century: The Causes and Consequences of Discrimination among Millennials in Two Experiments." Working paper.

Giulietti, Corrado, Mirco Tonin, and Michael Vlassopoulos. 2015. "Racial Discrimination in Local Public Services: A Field Experiment in the US." *IZA Discussion Paper* 9290 and *CESifo Working Paper* 5537. Retrieved June 13, 2016. http://ftp.iza.org/dp9290.pdf

Hanson, Andrew, and Zackary Hawley. 2011. "Do Landlords Discriminate in the Rental Housing Market? Evidence from an Internet Field Experiment in US Cities." *Journal of Urban Economics* 70(2-3):99-114. https://doi.org/10.1016/j.jue.2011.02.003

Hanson, Andrew, Zackary Hawley, Bo Liu, and Hal Martin. 2016. "Discrimination in Mortgage Lending: Evidence from a Correspondence Experiment." *Journal of Urban Economics* 92(2):48-65. https://doi.org/10.1016/j.jue.2015.12.004

Heckman, James J. 1998. "Detecting Discrimination." *Journal of Economic Perspectives* 12(2):101-16. https://doi.org/10.1257/jep.12.2.101

Heckman, James J., and Peter Siegelman. 1993. "The Urban Institute Audit Studies: Their Methods and Findings." Pp. 187–258 in *Clear and Convincing Evidence: Measurement of Discrimination in America*, edited by Michael Fix and Raymond. J. Stucyk. Lanham, MD: Urban Institute Press.

Hogan, Bernie, and Brent Berry. 2011. "Racial and Ethnic Biases in Rental Housing: An Audit Study of Online Apartment Listings." *City and Community* 10(4):351-72. https://doi.org/10.1111/j.1540-6040.2011.01376.x

Horne, Christine, Brice Darras, Elyse Bean, Anurag Srivastava, and Scott Firckel. 2015. "Privacy, Technology, and Norms: The Case of Smart Meters." *Social Science Research* 51:64-76. https://doi.org/10.1016/j.ssresearch.2014.12.003

Horton, John J., David G. Rand, and Richard J. Zeckhauser. 2011. "The Online Laboratory: Conducting Experiments in a Real Labor Market." *Experimental Economics* 14(3): 399-425. https://doi.org/10.1007/s10683-011-9273-9

Ipeirotis, Panagiotis. 2010. "Demographics of Mechanical Turk." NYU Working Paper Series.

Jacquemet, Nicolas, and Constantine Yannelis. 2012. "Indiscriminate Discrimination: A Correspondence Test for Ethnic Homophily in the Chicago Labor Market." *Labour Economics* 19(6):824-32. https://doi.org/10.1016/j.labeco.2012.08.004

Kirschenman, Joleen, and Kathryn M. Neckerman. 1991. "'We'd Love to Hire Them But...': The Meaning of Race to Employers." Pp. 203–32 in *The Urban Underclass*, edited by Christopher Jencks and Paul E. Peterson. Washington, D.C.: Brookings.

Levitt, Steven D., and Stephen J. Dubner. 2005. *Freakonomics: A Rogue Economist Explores the Hidden Side of Everything*. William Morrow.

Lieberson, Stanley. 2000. *A Matter of Taste: How Names, Fashions, and Culture Change.* New Haven, CT: Yale University Press.

Lieberson, Stanley, and Kelly S. Mikelson. 1995. "Distinctive African American Names: An Experimental, Historical, and Linguistic Analysis of Innovation." *American Sociological Review* 60(6):928-46. https://doi.org/10.2307/2096433

Mason, Winter, and Siddharth Suri. 2012. "Conducting Behavioral Research on Amazon's Mechanical Turk." *Behavior Research Methods* 44(1):1-23. https://doi.org/10.3758/s13428-011-0124-6

Milkman, Katherine L., Modupe Akinola, and Dolly Chugh. 2012. "Temporal Distance and Discrimination: An Audit Study in Academia." *Psychological Science* 23(7):710-7. https://doi.org/10.1177/0956797611434539

Mullinix, Kevin, Thomas J. Leeper, James Druckman, and Jeremy Freese. 2016. "The Generalizability of Survey Experiments." *Journal of Experimental Political Science* 2(2):109-138. https://doi.org/10.1017/XPS.2015.19

Paolacci, Gabriele, Jesse Chandler, and Panagiotis Ipeirotis. 2010. "Running Experiments on Amazon Mechanical Turk." *Judgment and Decision Making* 5(5):411-9.

Peer, Eyal, Joachim Vosgerau, and Alessandro Acquisti. 2014. "Reputation as a Sufficient Condition for Data Quality on Amazon Mechanical Turk." *Behavior Research Methods* 46(4):1023-31. https://doi.org/10.3758/s13428-013-0434-y

Rand, David G. 2012. "The Promise of Mechanical Turk: How Online Labor Markets Can Help Theorists Run Behavioral Experiments." *Journal of Theoretical Biology* 299: 172-9. https://doi.org/10.1016/j.jtbi.2011.03.004

Sears, David O. 1986. "College Sophomores in the Laboratory: Influences of a Narrow Data Base on Social Psychology's View of Human Nature." *Journal of Personality and Social Psychology* 51(3):515-30. https://doi.org/10.1037/0022-3514.51.3.515

Sharma, Rajiv, Arnab Mitra, and Miron Stano. 2015. "Insurance, Race/Ethnicity, and Sex in the Search for a New Physician." *Economics Letters* 137:150-3. https://doi.org/10.1016/j.econlet.2015.11.005

U.S. Census Bureau. 2012. "Genealogy Data: Frequently Occurring Surnames from Census 2000, Census Report Data File A: Top 1000 Names." Retrieved October 20, 2013. https://www.census.gov/topics/population/genealogy/data/2000_surnames.html

Weinberg, Jill, Jeremy Freese, and David McElhattan. 2014. "Comparing Data Characteristics and Results of an Online Factorial Survey between a Population-Based and Crowdsource-Recruited Sample." *Sociological Science* 1:292-310. https://doi.org/10.15195/v1.a19

Acknowledgements: An earlier version of this article was presented at the 2015 annual meeting of the American Sociological Association in Chicago, IL. Larry D. Schoen provided access to birth record data from New York. Anup Das, Qing Zheng, Betsy Cliff, and Neala Berkowski served as excellent research assistants on this project. I also thank Shawn Bauldry, Colleen Carey, Philip Cohen, Jonathan Daw, René Flores, Devah Pager, Lincoln Quillian, Charles Seguin, and Ashton Verdery for their helpful comments.

S. Michael Gaddis: Department of Sociology, University of California, Los Angeles.
E-mail: mgaddis@soc.ucla.edu.

E-lit-002: Racially Distinctive Names Signal Both Race/Ethnicity and Social Class

Charles Crabtree, S. Michael Gaddis, John B. Holbein and Edvard Nergard Larsen | Sociological Science 9: 454-472 | 2022-12-12

https://sociologicalscience.com/download/vol-9/december/SocSci_v9_454to472.pdf

About this source

The paper tests, rather than assumes, the excludability condition Bertrand and Mullainathan relied on - that Emily/Greg versus Lakisha/Jamal differ only in perceived race. It fails. Across three studies of education and income perceptions and two of social-class perceptions, respondents rank named individuals in a racially tiered order: Asian and White names at the top (both perceived as holding at least a bachelor's degree and roughly middle income), then Black names, then Hispanic names at the bottom on every dimension. So a canonical name treatment is double-barreled, and part of what an audit measures as race is a class inference. Two results cut the other way, against reading the canonical estimate as mostly a class artifact. First, class perceptions do vary within racial groups - a few White names read as below middle class and a few Black names as above it - yet the vast majority of the White distribution does not overlap the vast majority of the Black distribution, so names cannot be matched on perceived class without leaving the racially distinctive pool. Second, and more directly, Study 5 pins education down explicitly: respondents told that a named person holds a high school degree, a state-school bachelor's degree or an Ivy League bachelor's degree still place Black-named individuals in a lower social class than White-named individuals at every level; the gap narrows significantly at the Ivy League level but remains significant at $p < 0.05$, which the authors gloss as requiring 'incredibly elite' education for a Black-named person to be seen as White-named peers are. Individual names matter too: Bertrand and Mullainathan's own Latonya and Tanisha are among seven first names whose perceived education differs significantly from their group mean, and Lakisha is among five for income. The authors decline to rule on whether class signaling is a bug or a feature: if class inference is a channel through which racial discrimination operates, netting it out would be post-treatment bias.

Bound as the original PDF. Retrieved 2026-09-10. sha256 5ad9c1fc980d5a69...

Bound into an Economic Briefing · by Claude, may contain mistakes

Full document begins on the next page.

E-lit-002: https://sociologicalscience.com/download/vol-9/december/SocSci_v9_454to472.pdf

sociological science

Racially Distinctive Names Signal Both Race/Ethnicity and Social Class

Charles Crabtree,[a] S. Michael Gaddis,[b] John B. Holbein,[c] Edvard Nergård Larsen[d]

a) Dartmouth College; b) NWEA and University of California, Los Angeles; c) University of Virginia; d) University of Oslo

Abstract: Researchers studying discrimination and bias frequently conduct experiments that use racially distinctive names to signal race or ethnicity. The evidence that these studies provide about racial discrimination depends on the assumption that the names researchers use differ only based on perceived race and not some other factor. In this article, we assess this common assumption using data from five different studies (n = 1,004; 2,002; 1,035; 5,631; 1,858) conducted at different times across four separate survey platforms (Lucid Marketplace, Lucid Theorem, MTurk, and Prolific). We find evidence that names commonly used to signal race/ethnicity also influence perceptions about socioeconomic status and social class. Specifically, we observe that Americans tend to think that individuals with names typically used by Black and Hispanic people have lower educational attainment and income and are of a lower social class. Even when we present respondents with the educational attainment of a named individual, respondents still perceive Black people as lower social class than White people. We discuss the implications of these findings for past and future experimental work that uses names to signal race. We also articulate the importance of choosing names that best approximate the quantity that scholars want to estimate.

Keywords: names; discrimination; race/ethnicity; socioeconomic status; social class; experiments

Citation: Crabtree, Charles, S. Michael Gaddis, John B. Holbein, and Edvard Nergård Larsen. 2022. "Racially Distinctive Names Signal Both Race/Ethnicity and Social Class." Sociological Science 9: 454-472.

Received: December 4, 2021

Accepted: February 21, 2022

Published: December 12, 2022

Editor(s): Jesper Sørensen, Jeremy Freese

DOI: 10.15195/v9.a18

Copyright: © 2022 The Author(s). This open-access article has been published under a Creative Commons Attribution License, which allows unrestricted use, distribution and reproduction, in any form, as long as the original author and source have been credited.

RESEARCHERS studying discrimination and bias frequently conduct experiments that use racially distinctive names to signal race or ethnicity.[1] This practice is perhaps most common in field experiments known as audit studies or correspondence audits, which provide the bulk of modern causal evidence on racial discrimination. Since the 1960s, researchers have conducted hundreds of audits to examine discrimination more broadly in various markets (Crabtree 2018; Daniel 1968; Fix and Struyk 1993; Gaddis 2018b; Wienk et al. 1979). This body of research has found consistent evidence of racial discrimination in the labor and housing markets over the last six decades (Bertrand and Mullainathan 2004; Gaddis 2015; Gaddis and Ghoshal 2015, 2020; Gaddis et al. 2022b). More recently, researchers have expanded the use of audits to examine discrimination among a broader set of actors, such as bureaucrats, college admissions counselors, health care professionals, politicians, principals, professors, and members of the general public (Block et al. 2021; Butler and Broockman 2011; Einstein and Glick 2017; Gaddis et al. 2021; Gell-Redman et al. 2018; Hughes et al. 2020; Milkman, Akinola, and Chugh 2012; Pfaff et al. 2021). The popularity and usefulness of this method are reflected in numerous reviews (Baert 2018; Butler and Crabtree 2021; Gaddis and DiRago 2021), meta-analyses (Gaddis et al. 2022b; Quillian et al. 2017; Zschirnt and Ruedin 2016), and methodological examinations (Gaddis 2018a; Larsen 2020; Vuolo, Uggen, and

Lageson 2018) in the last few years alone. One reason for this popularity is that audits provide strong causal evidence of discrimination. They allow researchers to (1) avoid the typical shortcomings of using secondary data to examine discrimination (Crabtree 2019; Golder, Crabtree, and Dhima 2019) and (2) bypass the possibility of social desirability bias from respondents in surveys and interviews (Butler and Crabtree 2021; Gaddis 2019, 2022).

Correspondence audits, however, are not alone in their use of names for studying racial discrimination. Researchers also often use conjoint and survey experiments to uncover the natures and origins of discrimination. Like correspondence audits, these experiments have exploded in popularity in recent years (Crabtree et al. 2021; Druckman and Green 2021; Hainmueller, Hangartner, and Yamamoto 2015; Hainmueller and Hopkins 2015; Jenke et al. 2019; Lu et al. 2021). Although scholars use conjoint and survey experiments for many purposes, one of their more common uses is to study racial attitudes and preferences that are likely the foundation of discriminatory behavior (Caruso, Rahnev, and Banaji 2009; Doherty, Dowling, and Miller 2019; Jankowski, Prokop, and Tepe 2020). Research findings show that respondent preferences in conjoint experiments (and discrete choice experiments specifically) generalize reasonably well to real-world behavior (Hainmueller et al. 2015).

Many correspondence audits and conjoint experiments, however, do not signal race directly but rather indirectly through the use of names. The indirect signaling of race is potentially a problem for the interpretation of these experimental results because names might signal multiple characteristics in addition to race. One way of viewing this is as a possible violation of the excludability assumption (Gerber and Green 2012). In this case, the causal effect identified in a study that uses names might be a combination of the intended treatment (i.e., race) and other treatment aspects (i.e., social class) (Hansen and Tummers 2020). Another way of viewing this is as a violation of "information equivalence" (Dafoe, Zhang, and Caughey 2018) or "information leakage" (Sher and McKenzie 2006), which can occur when researchers present experimental subjects with information about one attribute in an experimental treatment (i.e., race) and this causes them to update their beliefs about others too (i.e., social class). A third way of viewing this is as a potential threat to the "construct validity" (Cronbach and Meehl 1955; Trochim 2001) of names as "race" treatments. This would mean that name treatments might not be adequate operationalizations of the underlying theoretical concept (i.e., race) because they also might measure other concepts (i.e., social class). In other words, the experimental treatments might feature a "double stimulus" (Converse and Presser 1986) or contain a "double-barreled treatment" (Kertzer and Brutger 2016).

Surprisingly, few studies empirically examine the construct validity of using names as measures to signal race or racial identity. Of those, many focus on the appropriateness of the race signal without deeply investigating whether names send signals *other* than race (Crabtree and Chykina 2018; Gaddis 2017a,b; Gaddis, Kreisberg, and Crabtree 2022a). This work suggests heterogeneity in how accurately and consistently individuals perceive the race that researchers intend to signal with names. Others have used demographic data on names and socioeconomic status (SES) to suggest that names likely only signal race and not SES or social

class (Bertrand and Mullainathan 2004; Fryer and Levitt 2004) or that, at the very least, experimental subjects are not influenced by any signals of SES (Butler and Homola 2017; Carnes and Holbein 2019).[2] However, a potential problem with social class signaling is that people often misperceive economic reality. This misperception is particularly true in the American context, where most people have little knowledge about economic conditions (Callaghan et al. 2021), the media provides disproportionate coverage of Black poverty (Roberts et al. 2019), and rich and poor individuals hold different views about the state of economic equality across racial groups (Kraus, Rucker, and Richeson 2017).

Two recent studies examined respondents' perceptions of social class using names (Barlow and Lahey 2018; Landgrave and Weller 2022). Although these studies move the literature forward in multiple ways, they still leave many important questions unanswered. In the first study, the authors use a very small sample size of undergraduate students (Barlow and Lahey 2018), which may not capture the heterogeneity in perceptions found in larger, more diverse samples (Crabtree and Chykina 2018; Gaddis 2017a,b). In the second study, the authors use a moderate sample size of respondents through Amazon's Mechanical Turk (MTurk). Still, they only examine respondents' perceptions of 12 total names—all male, four White, four Black, and four Hispanic (Landgrave and Weller 2022). Thus, the current literature in this area has two important limitations. First, Americans' misperceptions of racial and economic realities likely limit the value of using demographic data on naming patterns to proxy for perceptions. Second, existing work on perceptions does not have adequate coverage of a wide variety of names, examine nationally representative samples of respondents, or test multiple operationalizations and components of SES and social class. For these reasons, we believe that additional research in this area will help strengthen the construct validity of future experimental work using names to signal race.

In this article, we report the results of five studies we conducted to examine respondent perceptions of individuals' educational attainment, income, and social class based on racialized names. In total, 11,530 respondents provided 151,899 perceptions on 1,000 different combinations of first and last names. Using data from four of these studies, we find that respondents perceive educational attainment, income, and social class from names in a racially tiered pattern: White and Asian people are perceived at the top of the social class hierarchy, followed by Black people and then Hispanic people. Using data from the fifth study, we provide respondents with a vignette describing the named person as having attained either a high school degree, a bachelor's degree from a public state school, or a bachelor's degree from a private Ivy League school. Even when presented with an explicit signal of educational attainment, respondents perceived individuals with names commonly used by Black people as of a lower social class than individuals with names commonly used by White people. This gap in perceptions, however, is much smaller when names are coupled with a signal of high SES, in this case being described as having a bachelor's degree from a private Ivy League school. We discuss the implications that these results have for past and future experimental work that uses names to understand racial discrimination and outline several paths

scholars might take to understand more about the signaling power of names with regard to racial/ethnic discrimination and bias.

Data

We assess the construct validity of racialized names as an indirect signal of race using data from five different studies (n = 1,004; 2,002; 1,035; 5,631; 1,858) conducted at different times across four separate survey platforms (Lucid Marketplace, Lucid Theorem, MTurk, and Prolific). Our data focus on name evaluations, or perceptions of individuals based on their names. We collected survey data from U.S. residents on name evaluations across five separate data collections.

We conducted the first survey (hereafter Study 1) with 1,004 respondents via Prolific, an increasingly popular survey firm among social scientists (Palan and Schitter 2018), on February 4, 2021. We conducted the second survey (Study 2) with a national, quota-based (gender, age, region, race, ethnicity) sample of 2,002 Americans through Lucid Marketplace, an established and widely used survey provider (Coppock and McClellan 2019), from May 19 to 31, 2021. We conducted the third survey (Study 3) with a quota-based (gender, age, region) sample of 1,035 Americans who identify as first- or second-generation immigrants through Lucid Marketplace from May 18 to 27, 2021. We conducted the fourth survey (hereafter Study 4) with 5,631 respondents via MTurk from September 2014 through August 2015. We conducted the fifth survey (hereafter Study 5) with 1,858 respondents via MTurk in October 2015.[3]

In Studies 1 through 3, we asked respondents to answer multiple questions about their perceptions of each name. These perceptions span four different dimensions: race, citizenship status, income, and education.[4] In Studies 4 and 5, we only asked respondents about their social class perception of each name. We provide the specific questions about name perceptions in the "Dependent Variables" section below. Additionally, we asked respondents to report on a series of demographic and background questions in Studies 2 through 5. In Studies 2 and 3, we asked respondents to report their age, gender, income, educational attainment, race/ethnicity, and whether they are an American citizen. In Studies 4 and 5, we asked respondents to report their age, gender, household income, educational attainment, race/ethnicity, sex, relationship status, employment status, zip code, and whether they have any children younger than 18.

In sum, there are at least seven important dimensions of variation across our five studies: (1) respondent recruitment platform; (2) time period; (3) sample size; (4) specific names tested, as we describe below; (5) questions about social class; (6) answer options for social class questions; and (7) whether we provided an additional signal of SES to respondents.

Names

In Studies 1 through 3, we used Tzioumis (2018) for our primary list of first names and Comenetz (2016) for our list of last names. These are perhaps the two most

authoritative lists of name prevalence in the United States containing racial associations. These data sets allow us to determine the 10 most frequently used, racially distinctive (or differentially expressed) first names and 10 most frequently used, racially distinctive last names among White, Black or African American, Asian or Pacific Islander, and Hispanic people. First, we subset the first name list to those that are at least 90 percent used by Asian or Pacific Islander, Hispanic, and White people and at least 80 percent used by Black or African Americans.[5] We then subset the last name list to include those that are at least 90 percent used by Asian or Pacific Islander and White people, at least 80 percent used by Hispanic people, and at least 45 percent used by Black or African American people.[6] We then created all permutations of first and last names for each racial condition, giving us 400 distinct names ($10 \times 10 \times 4$).

In Studies 4 and 5, we selected first names used in prior experiments and then others from the Levitt and Dubner (2005) names list. Additionally, we selected last names using frequently occurring surnames from the 2000 Census, which lists the population racial composition of last names in the United States (U.S. Census Bureau 2012). In Study 4, we selected 80 White, 80 Black, and 40 Hispanic first names each paired with three randomly selected last names from the top six racially distinctive frequently occurring surnames list. Thus, we examine 600 total names: 240 White, 240 Black, and 120 Hispanic names. In Study 5, we selected a subset (from Study 4) of 80 total names: 40 White and 40 Black names.

Dependent Variables

In Studies 1 through 3, we examine respondents' perceptions of educational attainment and income based on names. One question asks respondents "What would be the highest level of education you think a person with this name has likely completed?" with the multiple-choice answer options "High school," "Bachelor's degree," "Master's degree," and "Ph.D. degree." Another asks respondents "What type of income do you think a person with this name would make?" with the multiple-choice answer options "Low income (Less than $41,000)," "Middle income ($41,000–$120,400)," and "Upper income (More than $120,400)." In Studies 4 and 5, we examine perceptions of social class. The survey question in Study 4 instructs respondents, "For each of the following names, list the social class category that you associate with that name (for example: lower class, working class, middle class, upper class, etc.). If you do not have a clear social class association with a name, you may type 'none.' " We recoded all answers into three categories: "lower/working class," "middle class," and "upper class." We removed "none" and blank responses from the analysis sample (9.7 percent). In Study 5, respondents are presented with a single vignette stating that "[Name] has a [degree]," where [Name] was one of 40 randomly assigned names (20 White and 20 Black) and [degree] was one of three randomly assigned educational attainment statuses: "high school degree," "bachelor's (college) degree from a public state school," or "bachelor's (college) degree from a private Ivy League school." The survey question in Study 5 asks respondents "What social class do you think [Name] is?" with the multiple-choice answer options "Lower or Working," "Middle," and "Upper."

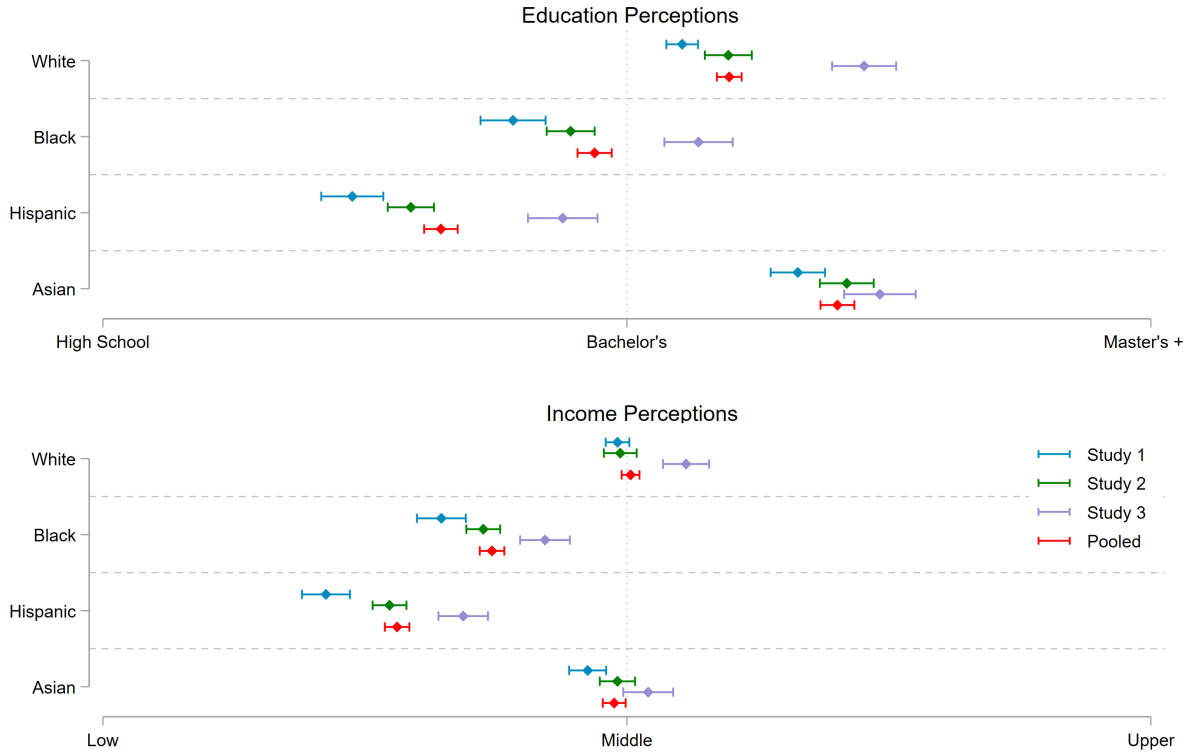


Figure 1: Education and income perceptions by name-signaled race. *Note:* Data from Studies 1 through 3. This plot shows respondent perceptions of educational attainment and income based on names intended to signal different racial/ethnic categories. Plotted points denote predictions and bars represent 95 percent confidence intervals.

Analysis and Results

We use these respondent-evaluation data to examine the construct validity of names as "race" treatments. To do this, we estimate a series of ordinary least squares models that regress respondent answers to our social class questions on a set of dummy variables that indicate the racial signal each name was intended to convey. In the figures below, we present predicted social class perceptions across racial groups. Plotted points denote predictions and bars represent 95 percent confidence intervals calculated using robust standard errors. We present the results from Studies 1 through 3 and a pooled analysis in Figure 1 with education perceptions in the top panel and income perceptions in the bottom panel. We examine heterogeneity by individual names in Figure 2. We present the results on social class perceptions from Study 4 in Figures 3 and 4. Finally, we present the results on social class perceptions with described educational attainment from Study 5 in Figure 5.

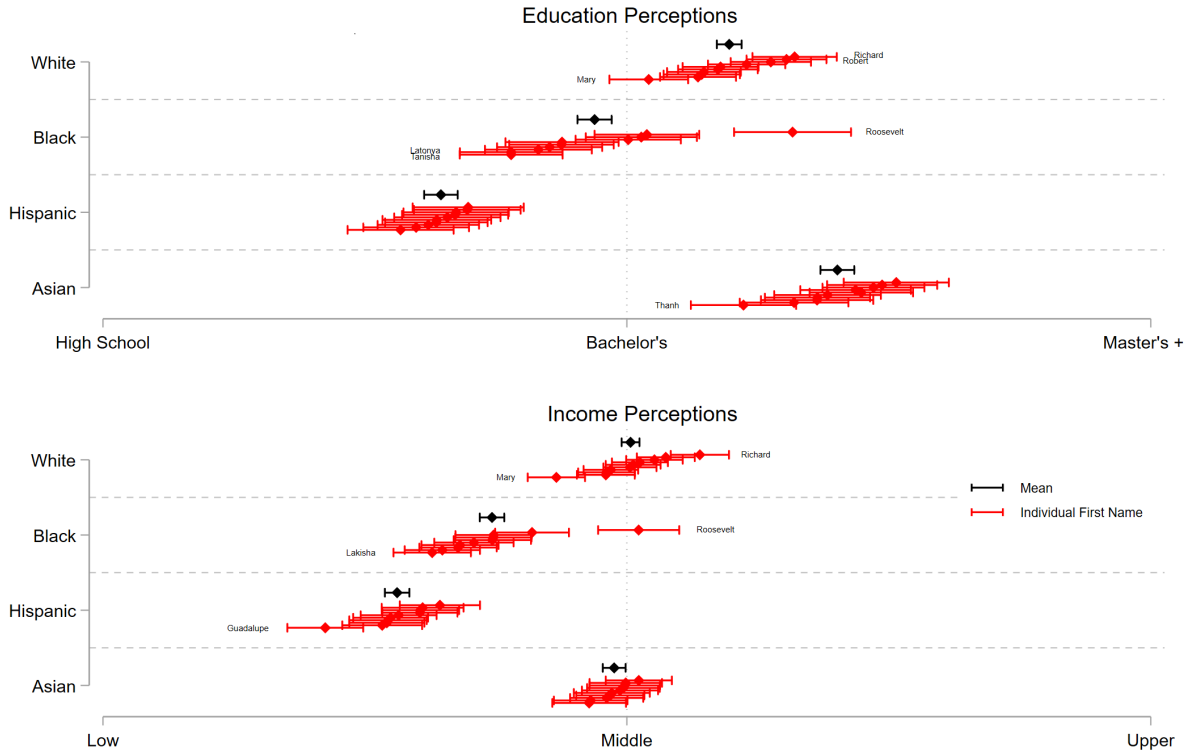


Figure 2: Heterogeneity in education and income perceptions by first names. *Note:* Pooled data from Studies 1 through 3 shown. This plot shows respondent perceptions of educational attainment and income based on names intended to signal different racial/ethnic categories. Plotted points denote predictions and bars represent 95 percent confidence intervals.

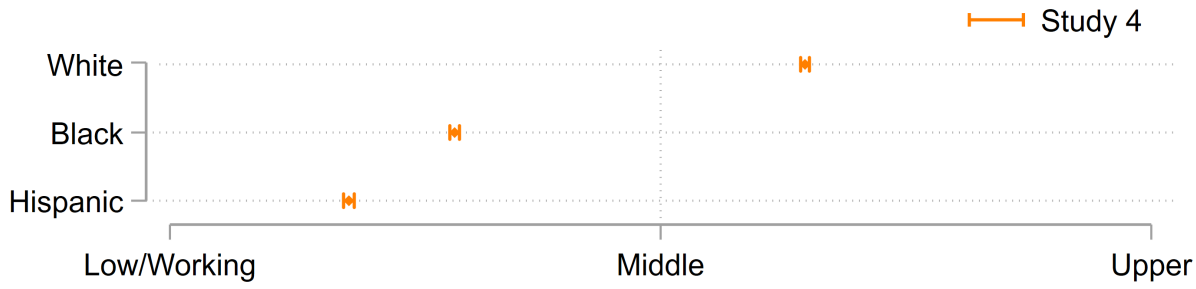


Figure 3: Social class perceptions by name-signaled race. *Note:* Data from Study 4. This plot shows respondent perceptions of social class based on names intended to signal different racial/ethnic categories. Plotted points denote predictions and bars represent 95 percent confidence intervals.

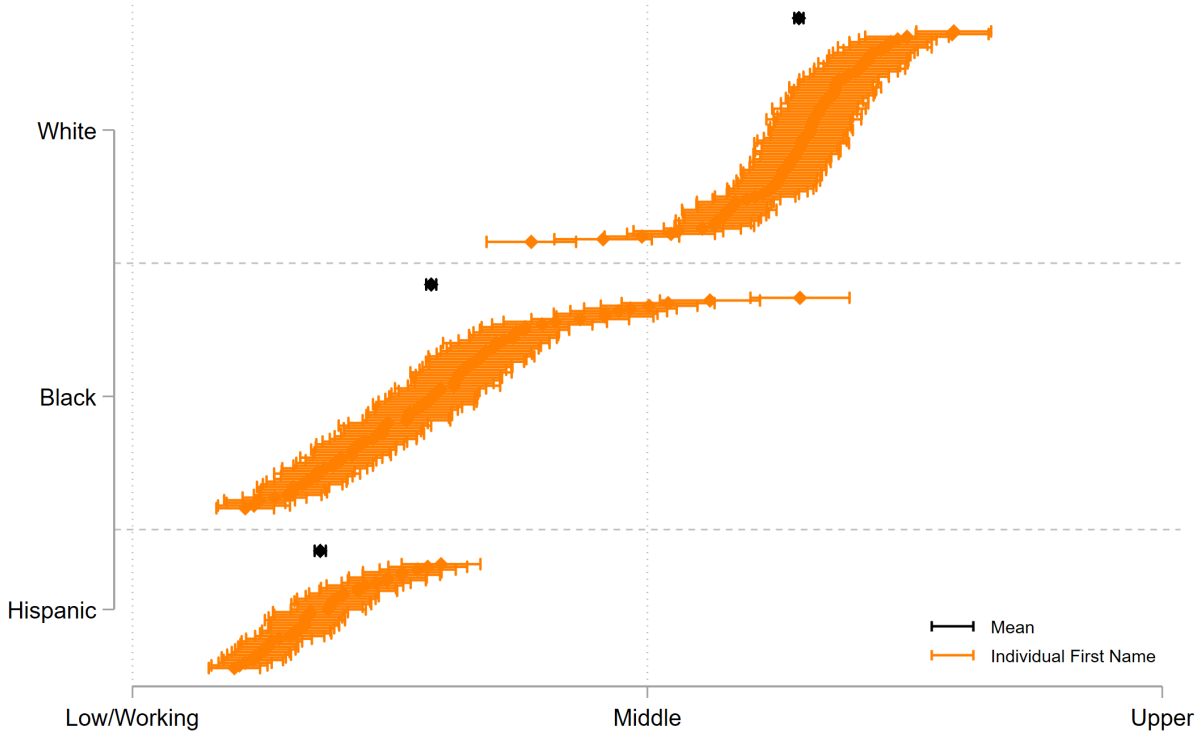


Figure 4: Heterogeneity in social class perceptions by first names. *Note:* Data from Study 4. This plot shows respondent perceptions of social class based on names intended to signal different racial/ethnic categories. Plotted points denote predictions and bars represent 95 percent confidence intervals.

In the top panel of Figure 1, we show that, on average, respondents view individuals with Asian names as having slightly higher educational attainment than individuals with White names. Both groups are perceived as having at least a bachelor's degree. Respondents view individuals with Black names as having lower educational attainment than individuals with White and Asian names, followed by individuals with Hispanic names. These findings are robust across the three studies, with some small differences in Study 3.

In the bottom panel of Figure 1, we show that, on average, respondents view individuals with White and Asian names as being very similar in terms of income—roughly the middle class. Similarly to the educational attainment results, respondents view individuals with Black names as having lower income than individuals with White and Asian names, followed by individuals with Hispanic names. These findings are also robust across the three studies.

In Figure 2, we show heterogeneity in education and income perceptions by each of the individual first names we used. The black bars represent 95 percent confidence intervals for each group, and the red bars represent 95 percent confidence intervals for each name. We include text for first names significantly different from their racial/ethnic group mean. This difference occurs in seven cases for education perceptions: Richard, Robert, Mary (White names); Roosevelt, Latonya, and Tanisha

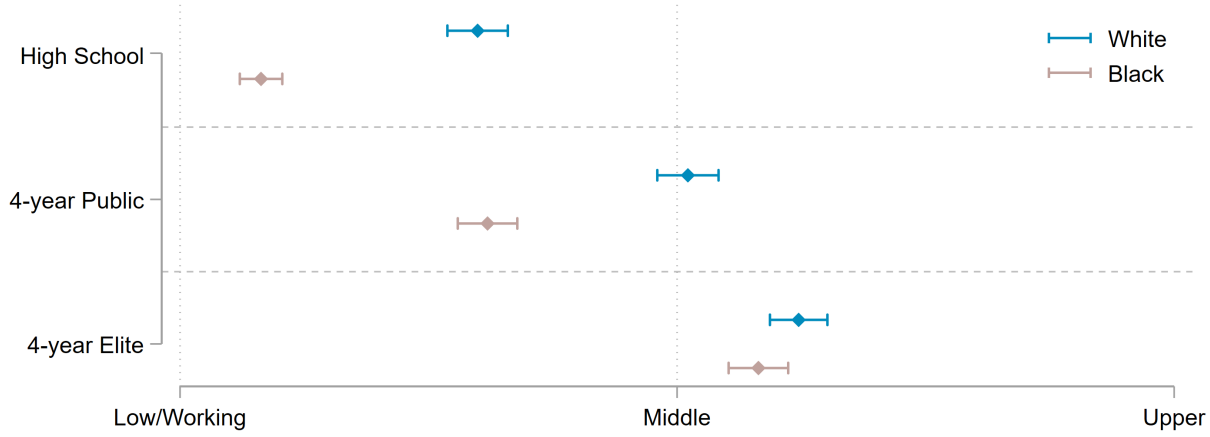


Figure 5: Social class perceptions by name-signaled race and described educational attainment. *Note:* Data from Study 5. This plot shows respondent perceptions of social class based on names intended to signal different racial/ethnic categories and described educational attainment. Plotted points denote predictions and bars represent 95 percent confidence intervals.

(Black names); and Thanh (Asian name). This difference occurs in five cases for income perceptions: Richard and Mary (White names); Roosevelt and Lakisha (Black names); and Guadalupe (Hispanic name).

In Figure 3, we show the group means of social class perceptions from Study 4. Individuals with White names are perceived as falling somewhere between middle and upper class. Individuals with Black names, on the other hand, are perceived as between middle and lower or working class, and individuals with Hispanic names slightly below that.

In Figure 4, we show heterogeneity in social class perceptions by each of the individual first names we used. The black bars represent 95 percent confidence intervals for each group, and the orange bars represent 95 percent confidence intervals for each name. We examine more names in Study 4; thus, we do not highlight specific names. However, the distribution shows that respondents perceive names across a wide range of social class categories for each racial/ethnic group. Respondents perceive a few White names as less than middle class, on average, and a few Black names as greater than middle class. However, the vast majority of the White distribution does not overlap with the vast majority of the Black distribution. Additionally, respondents' perceptions of social class for Hispanic names are clustered around the lower tail of the Black name distribution.

Finally, in Figure 5, we show the results from Study 5, which includes a direct signal of educational attainment. Even when presented with information that a named individual has a high school degree or a bachelor's degree from a public state school, respondents perceive individuals with Black names as of a lower social class than individuals with White names. This perception gap between individuals with Black and White names is reduced considerably ($p < 0.05$) when respondents are told that a named individual has a bachelor's degree from a private Ivy League

school. However, the difference in perception between individuals with Black and White names is still statistically significant ($p < 0.05$). Figure 5 provides an important topical finding that merits further investigation: for respondents to perceive Black people as having (roughly) equal SES levels as White people, one has to signal that Black people have incredibly elite levels of education.

Conclusion

The results from our five studies provide strong evidence that names commonly used in experimental designs to measure discrimination convey information about race as well as education, income, and social class. In other words, signals of race covary with perceptions of social class. This simple insight raises substantial issues for experimentalists who want to use names to signal race or ethnicity. It also suggests that we should read the results of prior experimental work using names to signal race with some careful consideration. Although the results from these studies are often discussed as showing how racialized names lead to differential outcomes, our findings suggest that these documented disparities might be, to some degree, driven by perceptions of social class as well. This is particularly true in cases where researchers do not include independent cues about social class in their experimental designs, such as in most audit studies outside employment contexts.

Still, we urge some caution regarding the external validity of this research as it applies to field and conjoint experiments. The survey experiment context necessitates that we *prompt* respondents to think about their social class perceptions of names. Hiring managers evaluate resumes and landlords evaluate rental applications without anyone asking them to think specifically about names. Although it might make sense that participants in these fields would consider the SES of those they interact with, it is less clear that this is true in another potential study types, such as those that involve public officials or members of the public. Moreover, in some contexts, decision-makers may have access to explicit signals of social class. In Study 5, we attempted to approximate the real-world process more closely by including another signal of social class. However, future work might conduct additional survey experiments with decision-makers without directly asking them about these perceptions. Researchers could use other questions and combinations of signals to examine this issue in more detail.

Where should researchers go from here? We present two potential broader interpretations of our findings. These two interpretations depend on whether scholars see the correlation between race and social class perceptions as a mechanism of inequality or a methodological problem to solve.

First, some scholars might consider that names signaling both social class and race is a "feature, not a bug." In other words, part of the mechanism of racial/ethnic discrimination might be individuals' associations of lower SES or social class with Black and Hispanic people. In this scenario, attempting to control for or adjust the impact of social class signaling would be misguided. It could even potentially lead to something akin to posttreatment bias. Researchers could develop new name-based, factorial experimental designs that focus on identifying social class as a mechanism driving discrimination. These designs could assess the extent to

which any observed racial discrimination might be the result of social class by crossing racial treatments with treatments related to social class, such as education or income. Others have implemented similar designs to examine racial discrimination alongside signals of sexual orientation and unemployment histories (Pedulla 2014, 2018b). This type of design would allow researchers to examine whether discrimination against members of a racial group varies as a function of social class or, considered differently, how much of any racial effects might be driven by different perceptions of social class. As we show in our experimental results, the strong link between perceptions of race and social class can be somewhat countered by providing independent signals of social class under specific conditions.

Scholars could also gather more data about the possible causes of any discrimination in a study. Researchers might want to consider pairing their experimental interventions with surveys or interviews from a sample of their target population (Pedulla 2018a). In surveys, researchers could ask respondents open-ended questions about why they make certain choices. Respondent replies might help shed light on the mechanisms driving any discrimination that they exhibit and thus help researchers identify whether the discrimination they document might be driven by racial animus, social class bias, or something else entirely. In a similar way, researchers could also conduct interviews to document potential mechanisms (Hou, Liu, and Crabtree 2020). Moreover, the social class signal might explain variations in discrimination across contexts where social class is more or less important and signaled more or less explicitly (Gaddis et al. 2022b).

Second, other scholars might consider that names signaling both social class and race is a "bug, not a feature." In other words, names are poor proxies for the intended signal of race. In this scenario, researchers might see a need to find different names across racial/ethnic categories that signal similar levels of social class. Theoretically, researchers could then use these names to send unbundled treatments of race/ethnicity. These types of experiments could serve a dual purpose. They might help researchers identify heterogeneity in racial/ethnic discrimination at different levels of social class. They might also provide estimates of racial/ethnic discrimination using "crisp" signals of race/ethnicity. This "crispness," however, comes at the expense of the additional assumption that perceptions of race do not cause perceptions of class. Therefore, controlling for the latter does not eliminate the effect of the former.

To be clear, our recommendation in this article is *not* that researchers stop using names to signal race. After all, other alternatives for signaling race—such as language difference, photographs, or video- or audio-based manipulations—may be *more* bundled than name-based treatments. Rather, there are two broad lessons to draw, one theoretical and one practical. In practice, researchers using names to signal race (whether in audit, conjoint, or other experiments) should, as standard practice, pretest the names they are using to ensure that names signal what the researchers intend. Second, researchers should be clearer about what it is they are seeking to capture when they are studying racial discrimination. If they desire to estimate the joint effect of race *and* social class, they should use names that manipulate both of these things. If they desire to, instead, partial out the effect of social class from race, then they should choose names and make design choices

that ensure that their experimental manipulations are not signaling both social class and race. Ultimately, these decisions require better theoretical clarity as to what estimands scholars desire to capture and more transparency in the choices researchers make in their experimental design and why these choices are made.

We do not believe that either interpretation is definitively correct. The best option likely depends on the research method and the context under examination. Researchers should chose name treatments that make sense in their particular application. We fully expect that much scholarly output will follow this article that branches in multiple directions. We think that the link between perceptions of race and social class means that using names necessarily will be messy. However, using names as a proxy for race likely captures or mimics real-world processes in a generally appropriate fashion. Researchers should take seriously that names are, on their own, bundled (multi-barreled) treatments (Converse and Presser 1986). Practically speaking, this means that researchers should interpret results with this in mind, explicitly acknowledging and understanding that in many cases, they cannot necessarily pull apart the "bundle of sticks" (Sen and Wasow 2016) that constitute race. Evidence of discrimination may capture actions based on racial bias, social class bias, or, as seems most likely to us, both.

Notes

1 Throughout this article, we use "race" as shorthand for both race and ethnicity. Social scientists often use them as synonyms, even if they treat the two concepts as substantively different (Bonilla-Silva 1997; Cornell and Hartmann 2006; Feagin and Feagin 1993; Omi and Winant 1994). For a critique of this approach, see Wimmer (2008). Although Wimmer (2008) would suggest using ethnicity instead of race to describe this group identity category, we use race in our data collection process and throughout this article.

2 Hereafter, we use SES and social class interchangeably while acknowledging that the terms denote related but different concepts.

3 Because Studies 2 through 5 captured respondent information, we applied for and received Institutional Review Board approval for Studies 2 and 3 at Dartmouth College (STUDY00032300 and STUDY00032301) and Studies 4 and 5 at the University of Michigan (HUM00082300). Study 1 did not collect any information about respondents and was considered just a coding task.

4 We did not randomize the presentation order of the questions on different dimensions of perception. Respondents answered questions first on racial perceptions, followed by citizenship status, income, and finally, education perceptions. Our design may have influenced respondents' answers by asking about racial perceptions before asking about income and education perceptions. Although this non-randomized order is not ideal, we believe the results from Study 4 help adjudicate this issue somewhat. In Study 4, we only asked respondents about social class perceptions. The patterns of results from Studies 1 through 3 are similar to those from Study 4. Thus, we believe that the inclusion of other perception questions and the question order may not have strongly influenced the results from Studies 1 through 3. Scholars could test this more directly in future work by randomizing the order of perception questions.

5 We needed to use a lower threshold because there are fewer names that are differentially used by Black or African American people compared with other races.

6 Again, we needed to use a lower threshold because there are fewer names that are differentially used by Black or African American people compared with other races.

References

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*, edited by S. Michael Gaddis, pp. 63–77. Cham: Springer.

Barlow, M. Rose, and Joanna N. Lahey. 2018. "What Race Is Lacey? Intersecting Perceptions of Racial Minority Status and Social Class." *Social Science Quarterly* 99(5):1680–98. https://doi.org/10.1111/ssqu.12529.

Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable Than Lakisha and Jamal?" *American Economic Review* 94(4):991–1013. https://doi.org/10.1257/0002828042002561.

Block, Ray, Jr., Charles Crabtree, John B. Holbein, and J. Quin Monson. 2021. "Are Americans Less Likely to Reply to Emails from Black People Relative to White People?" *Proceedings of the National Academy of Sciences* 118(52):e2110347118. https://doi.org/10.1073/pnas.2110347118.

Bonilla-Silva, Eduardo. 1997. "Rethinking Racism: Toward a Structural Interpretation." *American Sociological Review* 62(3):465–80. https://doi.org/10.2307/2657316.

Butler, Daniel M., and David E. Broockman. 2011. "Do Politicians Racially Discriminate against Constituents?" *American Journal of Political Science* 55(3):463–77. https://doi.org/10.1111/j.1540-5907.2011.00515.x.

Butler, Daniel M., and Charles Crabtree. 2021. "Audit Studies in Political Science." In *Advances in Experimental Political Science*, edited by James N. Druckham and Donald P. Green, pp. 42–55. Cambridge University Press.

Butler, Daniel M., and Jonathan Homola. 2017. "An Empirical Justification for the Use of Racially Distinctive Names to Signal Race in Experiments." *Political Analysis* 25(1):122–30. https://doi.org/10.1017/pan.2016.15.

Callaghan, Bennett, Leilah Harouni, Cydney H. Dupree, Michael W. Kraus, and Jennifer A. Richeson. 2021. "Testing the Efficacy of Three Informational Interventions for Reducing Misperceptions of the Black–White Wealth Gap." *Proceedings of the National Academy of Sciences* 118(38):e2108875118. https://doi.org/10.1073/pnas.2108875118.

Carnes, Nicholas, and John Holbein. 2019. "Do Public Officials Exhibit Social Class Biases When They Handle Casework? Evidence from Multiple Correspondence Experiments." *PLoS One* 14(3):e0214244. https://doi.org/10.1371/journal.pone.0214244.

Caruso, Eugene M., Dobromir A. Rahnev, and Mahzarin R. Banaji. 2009. "Using Conjoint Analysis to Detect Discrimination: Revealing Covert Preferences from Overt Choices." *Social Cognition* 27(1):128–37. https://doi.org/10.1521/soco.2009.27.1.128.

Comenetz, Joshua. 2016. "Frequently Occurring Surnames in the 2010 Census." Technical report, U.S. Census Bureau.

Converse, Jean M., and Stanley Presser. 1986. *Survey Questions: Handcrafting the Standardized Questionnaire*, volume 63 of *Quantitative Applications in the Social Sciences*. Thousand Oaks, CA: SAGE Publications.

Coppock, Alexander, and Oliver A. McClellan. 2019. "Validating the Demographic, Political, Psychological, and Experimental Results Obtained from a New Source of Online Survey Respondents." *Research & Politics* 6(1). https://doi.org/10.1177/2053168018822174.

Cornell, Stephen, and Douglas Hartmann. 2006. *Ethnicity and Race: Making Identities in a Changing World*. Thousand Oaks, CA: SAGE Publications, 2nd edition.

Crabtree, Charles. 2018. "An Introduction to Conducting Email Audit Studies." In *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*, edited by S. Michael Gaddis, pp. 103–17. Cham: Springer.

Crabtree, Charles. 2019. *Measuring and Explaining Discrimination*. Ph.D. thesis, University of Michigan.

Crabtree, Charles, and Volha Chykina. 2018. "Last Name Selection in Audit Studies." *Sociological Science* 5:21–8. https://doi.org/10.15195/v5.a2.

Crabtree, Charles, S. Michael Gaddis, Edvard Nergard Larsen, and John H. Holbein. 2021. "Changes in the Amount and Type of Social Science Research on Discrimination: 1950 to 2015." Posted December 6, 2021, on SSRN. https://doi.org/10.2139/ssrn.3977169.

Cronbach, Lee J., and Paul E. Meehl. 1955. "Construct Validity in Psychological Tests." *Psychological Bulletin* 52(4):281–302. https://doi.org/10.1037/h0040957.

Dafoe, Allan, Baobao Zhang, and Devin Caughey. 2018. "Information Equivalence in Survey Experiments." *Political Analysis* 26(4):399–416. https://doi.org/10.1017/pan.2018.9.

Daniel, W. W. 1968. *Racial Discrimination in England: Based on the PEP Report*. London: Penguin Books.

Doherty, David, Conor M. Dowling, and Michael G. Miller. 2019. "Do Local Party Chairs Think Women and Minority Candidates Can Win? Evidence from a Conjoint Experiment." *Journal of Politics* 81(4):1282–97. https://doi.org/10.1086/704698.

Druckman, James N., and Donald P. Green. 2021. *Cambridge Handbook of Advances in Experimental Political Science*. New York: Cambridge University Press.

Einstein, Katherine Levine, and David M. Glick. 2017. "Does Race Affect Access to Government Services? An Experiment Exploring Street-Level Bureaucrats and Access to Public Housing." *American Journal of Political Science* 61(1):100–116. https://doi.org/10.1111/ajps.12252.

Feagin, Joe R., and Clairece Booher Feagin. 1993. *Racial and Ethnic Relations*. Hoboken, NJ: Prentice-Hall.

Fix, Michael E., and Raymond J. Struyk. 1993. *Clear and Convincing Evidence: Measurement of Discrimination in America*. Washington, DC: Urban Institute Press.

Fryer, Roland G., Jr., and Steven D. Levitt. 2004. "Understanding the Black–White Test Score Gap in the First Two Years of School." *Review of Economics and Statistics* 86(2):447–64. https://doi.org/10.1162/003465304323031049.

Gaddis, S. Michael. 2015. "Discrimination in the Credential Society: An Audit Study of Race and College Selectivity in the Labor Market." *Social Forces* 93(4):1451–79. https://doi.org/10.1093/sf/sou111.

Gaddis, S. Michael. 2017a. "How Black Are Lakisha and Jamal? Racial Perceptions from Names Used in Correspondence Audit Studies." *Sociological Science* 4:469–89. https://doi.org/10.15195/v4.a19.

Gaddis, S. Michael. 2017b. "Racial/Ethnic Perceptions from Hispanic Names." *Socius* 3. https://doi.org/10.1177/2378023117737193.

Gaddis, S. Michael. 2018a. *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*. Cham: Springer.

Gaddis, S. Michael. 2018b. "An Introduction to Audit Studies in the Social Sciences." In *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*, edited by S. Michael Gaddis, pp. 3–44. Cham: Springer.

Gaddis, S. Michael. 2019. "Understanding the 'How' and 'Why' Aspects of Racial/Ethnic Discrimination: A Multimethod Approach to Audit Studies." *Sociology of Race and Ethnicity* 5(4):443–55. https://doi.org/10.1177/2332649219870183.

Gaddis, S. Michael. 2022. "Cognitive-Behavioral Disconnect Bias in the Study of Racial/Ethnic Discrimination: A Method to Examine and Measure Respondents' Alignment of Predicted and Actual Behavior." Posted January 31, 2022, on SSRN. https://doi.org/10.2139/ssrn.4021118.

Gaddis, S. Michael, Charles Crabtree, John B. Holbein, and Steven Pfaff. 2021. "Chinese Americans Faced High Levels of Discrimination in Public Schools before the COVID-19 Pandemic." Working paper.

Gaddis, S. Michael, and Nicholas DiRago. 2021. "Audit Studies of Housing in the United States: Established, Emerging, and Future Research." Posted March 24, 2021, on SSRN; revised December 13, 2021. https://doi.org/10.2139/ssrn.3796335.

Gaddis, S. Michael, and Raj Ghoshal. 2015. "Arab American Housing Discrimination, Ethnic Competition, and the Contact Hypothesis." *The ANNALS of the American Academy of Political and Social Science* 660(1):282–99. https://doi.org/10.1177/0002716215580095.

Gaddis, S. Michael, and Raj Ghoshal. 2020. "Searching for a Roommate: A Correspondence Audit Examining Racial/Ethnic and Immigrant Discrimination among Millennials." *Socius* 6. https://doi.org/10.1177/2378023120972287.

Gaddis, S. Michael, A. Nicole Kreisberg, and Charles Crabtree. 2022a. "Assessing Immigrant Generation and Citizenship Status from Names." Posted March 4, 2019, on SSRN; revised October 3, 2022. https://doi.org/10.2139/ssrn.3022217.

Gaddis, S. Michael, Edvard Nergard Larsen, Charles Crabtree, and John H. Holbein. 2022b. "Discrimination against Black and Hispanic Americans Is Highest in Hiring and Housing Contexts: A Meta-analysis of Correspondence Audits." Posted January 31, 2022, on SSRN. https://doi.org/10.2139/ssrn.3975770.

Gell-Redman, Micah, Neil Visalvanich, Charles Crabtree, and Christopher J. Fariss. 2018. "It's All about Race: How State Legislators Respond to Immigrant Constituents." *Political Research Quarterly* 71(3):517–31. https://doi.org/10.1177/1065912917749322.

Gerber, Alan S., and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. New York: W. W. Norton & Company.

Golder, Sona N., Charles Crabtree, and Kostanca Dhima. 2019. "Legislative Representation and Gender (Bias)." *Political Science* 71(1):1–16. https://doi.org/10.1080/00323187.2019.1632151.

Hainmueller, Jens, Dominik Hangartner, and Teppei Yamamoto. 2015. "Validating Vignette and Conjoint Survey Experiments against Real-World Behavior." *Proceedings of the National Academy of Sciences* 112(8):2395–400. https://doi.org/10.1073/pnas.1416587112.

Hainmueller, Jens, and Daniel J. Hopkins. 2015. "The Hidden American Immigration Consensus: A Conjoint Analysis of Attitudes toward Immigrants." *American Journal of Political Science* 59(3):529–48. https://doi.org/10.1111/ajps.12138.

Hansen, Jesper Asring, and Lars Tummers. 2020. "A Systematic Review of Field Experiments in Public Administration." *Public Administration Review* 80(6):921–31. https://doi.org/10.1111/puar.13181.

Hou, Yue, Chuyu Liu, and Charles Crabtree. 2020. "Anti-Muslim Bias in the Chinese Labor Market." *Journal of Comparative Economics* 48(2):235–50. https://doi.org/10.1016/j.jce.2019.12.001.

Hughes, D. Alex, Micah Gell-Redman, Charles Crabtree, Natarajan Krishnaswami, Diana Rodenberger, and Guillermo Monge. 2020. "Persistent Bias among Local Election Officials." *Journal of Experimental Political Science* 7(3):179–87. https://doi.org/10.1017/XPS.2019.23.

Jankowski, Michael, Christine Prokop, and Markus Tepe. 2020. "Representative Bureaucracy and Public Hiring Preferences: Evidence from a Conjoint Experiment among German Municipal Civil Servants and Private Sector Employees." *Journal of Public Administration Research and Theory* 30(4):596–618. https://doi.org/10.1093/jopart/muaa012.

Jenke, Libby, Kirk Bansak, Jens Hainmueller, and Dominik Hangartner. 2019. "Using Eye-Tracking to Understand Decision-Making in Conjoint Experiments." *Political Analysis* 29(1):75–101. https://doi.org/10.1017/pan.2020.11.

Kertzer, Joshua D., and Ryan Brutger. 2016. "Decomposing Audience Costs: Bringing the Audience Back into Audience Cost Theory." *American Journal of Political Science* 60(1):234–49. https://doi.org/10.1111/ajps.12201.

Kraus, Michael W., Julian M. Rucker, and Jennifer A. Richeson. 2017. "Americans Misperceive Racial Economic Equality." *Proceedings of the National Academy of Sciences* 114(39):10324–31. https://doi.org/10.1073/pnas.1707719114.

Landgrave, Michelangelo, and Nicholas Weller. 2022. "Do Name-Based Treatments Violate Information Equivalence? Evidence from a Correspondence Audit Experiment." *Political Analysis* 30(1):142–48. https://doi.org/10.1017/pan.2020.52.

Larsen, Edvard N. 2020. "Induced Competition in Matched Correspondence Tests: Conceptual and Methodological Considerations." *Research in Social Stratification and Mobility* 65:100475. https://doi.org/10.1016/j.rssm.2020.100475.

Levitt, Steven D., and Stephen J. Dubner. 2005. *Freakonomics: A Rogue Economist Explores the Hidden Side of Everything*. New York: William Morrow.

Lu, Yao, Neeraj Kaushal, Xiaoning Huang, and S. Michael Gaddis. 2021. "Priming COVID-19 Salience Increases Prejudice and Discriminatory Intent against Asians and Hispanics." *Proceedings of the National Academy of Sciences* 118(36):e2105125118. https://doi.org/10.1073/pnas.2105125118.

Milkman, Katherine L., Modupe Akinola, and Dolly Chugh. 2012. "Temporal Distance and Discrimination: An Audit Study in Academia." *Psychological Science* 23(7):710–17. https://doi.org/10.1177/0956797611434539.

Omi, Michael, and Howard Winant. 1994. *Racial Formation in the United States: From the 1960s to the 1990s*. New York: Routledge, 2nd edition.

Palan, Stefan, and Christian Schitter. 2018. "Prolific.ac—A Subject Pool for Online Experiments." *Journal of Behavioral and Experimental Finance* 17:22–27. https://doi.org/10.1016/j.jbef.2017.12.004.

Pedulla, David S. 2014. "The Positive Consequences of Negative Stereotypes: Race, Sexual Orientation, and the Job Application Process." *Social Psychology Quarterly* 77(1):75–94. https://doi.org/10.1177/0190272513506229.

Pedulla, David S. 2018a. "Emerging Frontiers in Audit Study Research: Mechanisms, Variation, and Representativeness." In *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*, edited by S. Michael Gaddis, pp. 179–95. Cham: Springer.

Pedulla, David S. 2018b. "How Race and Unemployment Shape Labor Market Opportunities: Additive, Amplified, or Muted Effects?" *Social Forces* 96(4):1477–506. https://doi.org/10.1093/sf/soy002.

Pfaff, Steven, Charles Crabtree, Holger L. Kern, and John B. Holbein. 2021. "Do Street-Level Bureaucrats Discriminate Based on Religion? A Large-Scale Correspondence Experiment among American Public School Principals." *Public Administration Review* 81(2):244–59. https://doi.org/10.1111/puar.13235.

Quillian, Lincoln, Devah Pager, Ole Hexel, and Arnfinn 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):10870–75. https://doi.org/10.1073/pnas.1706255114.

Roberts, Caroline, Emily Gilbert, Nick Allum, and Léïla Eisner. 2019. "Research Synthesis: Satisficing in Surveys: A Systematic Review of the Literature." *Public Opinion Quarterly* 83(3):598–626. https://doi.org/10.1093/poq/nfz035.

Sen, Maya, and Omar Wasow. 2016. "Race as a Bundle of Sticks: Designs That Estimate Effects of Seemingly Immutable Characteristics." *Annual Review of Political Science* 19:499–522. https://doi.org/10.1146/annurev-polisci-032015-010015.

Sher, Shlomi, and Craig R. M. McKenzie. 2006. "Information Leakage from Logically Equivalent Frames." *Cognition* 101(3):467–94. https://doi.org/10.1016/j.cognition.2005.11.001.

Trochim, William M. K. 2001. *Research Methods Knowledge Base*. Cincinnati, OH: Atomic Dog Publishing, 2nd edition.

Tzioumis, Konstantinos. 2018. "Demographic Aspects of First Names." *Scientific Data* 5:180025. https://doi.org/10.1038/sdata.2018.25.

U.S. Census Bureau. 2012. "Frequently Occurring Surnames from Census 2000; Census Report Data File A: Top 1000 Names." Accessed October 20, 2013. https://www.census.gov/topics/population/genealogy/data/2000_surnames.html.

Vuolo, Mike, Christopher Uggen, and Sarah Lageson. 2018. "To Match or Not to Match? Statistical and Substantive Considerations in Audit Design and Analysis." In *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*, edited by S. Michael Gaddis, pp. 119–140. Cham: Springer.

Wienk, Ronald E., Clifford E. Reid, John C. Simonson, and Frederick J. Eggers. 1979. "Measuring Racial Discrimination in American Housing Markets: The Housing Market Practices Survey." Technical report, Department of Housing and Urban Development, Office of Policy Development and Research.

Wimmer, Andreas. 2008. "The Making and Unmaking of Ethnic Boundaries: A Multilevel Process Theory." *American Journal of Sociology* 113(4):970–1022. https://doi.org/10.1086/522803.

Zschirnt, Eva, and Didier Ruedin. 2016. "Ethnic Discrimination in Hiring Decisions: A Meta-analysis of Correspondence Tests 1990–2015." *Journal of Ethnic and Migration Studies* 42(7):1115–34. https://doi.org/10.1080/1369183X.2015.1133279.

Acknowledgments: We thank the service workers and small businesses in San Francisco's Mission District for the bountiful supply of burritos that provided fuel for the authors' intense writing retreat that resulted in this article. We also thank NBA League Pass.

Charles Crabtree: Department of Government, Dartmouth College. E-mail: crabtree@dartmouth.edu; URL: charlescrabtree.com.

S. Michael Gaddis: Senior Research Scientist, NWEA; Department of Sociology, University of California, Los Angeles; and California Center for Population Research. E-mail: mgaddis@soc.ucla.edu; URL: stevenmichaelgaddis.com.

John B. Holbein: Frank Batten School of Leadership and Public Policy, University of Virginia. E-mail: holbein@virginia.edu; URL: sites.google.com/site/johnbholbein/.

Edvard Nergård Larsen: Department of Sociology and Human Geography, University of Oslo. E-mail: e.n.larsen@sosgeo.uio.no; URL: sv.uio.no/iss/english/people/aca/edvardnl.

E-lit-003: The Causes and Consequences of Distinctively Black Names

Roland G. Fryer, Jr. and Steven D. Levitt | Quarterly Journal of Economics 119(3): 767-805 | 2004-08-01

library: fryer levitt_2004_distinctively black names_qje.pdf (Jason's library, published QJE version)

About this source

The paper's first finding is that Black and White naming converged in the early 1960s and then diverged sharply in a seven-year window in the early 1970s, with the median Black girl in a segregated neighborhood moving from a name twice as likely to be Black to one more than twenty times as likely; the mean Black Name Index rose from 60.9 in 1961-1967 to 71.0 in 1989-2000, and by the recent period more than 40 percent of Black girls received a name given to no White girl born in California that year. The second, and the one that matters for a hiring audit, is that this shift 'ghettoized' the names: until the late 1970s the Blackness of a name was only weakly correlated with parental characteristics, but in the 1980s and 1990s it became strongly associated with young, poor, unmarried and less-educated mothers, and the R-squared of the naming regressions climbs steadily across the four periods. The third is that, linking women born in 1973-1974 to their own births in 1989-2000, the coefficient on a woman's own Black Name Index is essentially zero across most adult outcomes once birth circumstances are controlled for: moving from a BNI of 50 to 100 shifts the per capita income of Blacks in her zip code by $100 and the probability of an unmarried birth by two-tenths of a percentage point, while the race dummy in the pooled specification is enormous (28.8 percentage points more Black in the neighborhood, over $2,000 lower zip code income). Fourth, and directly against the 'race as such' reading of the audit literature, Table IV shows the name is a genuine predictor of unobserved background even after conditioning on resume-like controls: a Black woman with a BNI of 100 is 20.9 percentage points more likely to have been born to a teenage mother and 31.3 points more likely to have been born out of wedlock than a Black woman in the same zip code with the same age and education whose name is equally common among Blacks and Whites - so the signal survives even inside an all-Black pool, though it is smaller there. The authors therefore offer three reconciliations with Bertrand and Mullainathan: names are a noisy race signal that stops mattering at the interview; names carry productivity-relevant information beyond the resume (statistical discrimination); or names have a small causal effect their coarse outcomes cannot detect. They explicitly decline to rule out employer

Bound as the original PDF. Retrieved 2026-09-10. sha256 356053d28ec90f84...

Bound into an Economic Briefing · by Claude, may contain mistakes

Full document begins on the next page.

E-lit-003: library: fryer levitt_2004_distinctively black names_qje.pdf (Jason's library, published QJE version)

THE
QUARTERLY JOURNAL
OF ECONOMICS

Vol. CXIX August 2004 Issue 3

THE CAUSES AND CONSEQUENCES OF DISTINCTIVELY BLACK NAMES*

ROLAND G. FRYER, JR. AND STEVEN D. LEVITT

In the 1960s Blacks and Whites chose relatively similar first names for their children. Over a short period of time in the early 1970s, that pattern changed dramatically with most Blacks (particularly those living in racially isolated neighborhoods) adopting increasingly distinctive names, but a subset of Blacks actually moving toward more assimilating names. The patterns in the data appear most consistent with a model in which the rise of the Black Power movement influenced how Blacks perceived their identities. Among Blacks born in the last two decades, names provide a strong signal of socioeconomic status, which was not previously the case. We find, however, no negative relationship between having a distinctively Black name and later life outcomes after controlling for a child's circumstances at birth.

I. INTRODUCTION

On May 17, 1954, the landmark Supreme Court decision in *Brown v. Board of Education of Topeka, Kansas* ruled, unanimously, that segregation in public schools was unconstitutional.

* We are grateful to George Akerlof, Susan Athey, Gary Becker, David Card, John Donohue, Esther Duflo, Edward Glaeser, Shoshana Grossbard-Shechtman, Daniel Hamermesh, Lawrence Katz, Glenn Loury, Douglas McAdam, Andrei Shleifer, and Steven Tadelis for helpful comments and suggestions. We thank seminar participants too numerous to mention for their comments and suggestions. We also thank two editors and three anonymous referees for detailed feedback that improved the paper. Michael Roh provided excellent research assistance. Financial support was provided by the National Science Foundation. A portion of this paper was written while Fryer was an NSF Post-Doctoral Fellow at the University of Chicago and Levitt was a fellow at the Center for Advanced Study in the Behavioral Sciences in Stanford. Correspondence can be addressed to Fryer at Harvard University, Littauer Center, Cambridge MA 02138 (e-mail: rfryer@fas.harvard.edu); or Levitt at Department of Economics, University of Chicago, 1126 East 59th Street, Chicago, IL 60637 (e-mail: slevitt@midway.uchicago.edu).

© 2004 by the President and Fellows of Harvard College and the Massachusetts Institute of Technology.
The Quarterly Journal of Economics, August 2004

This ruling paved the way for the fall of Jim Crow and large-scale desegregation. In the 1960s a series of further government actions were taken with the goal of achieving racial equality and integration, most notably the Civil Rights Act of 1964, Executive Order 11246 in 1965, and the Fair Housing Act of 1968. The civil rights movement arguably represents one of the most profound social transformations in American history [Woodward 1974; Young 1996].

Nonetheless, an enormous racial divide persists. There are large disparities between Blacks and Whites in the United States on many indicators of social and economic welfare including income [Bound and Freeman 1992; Chandra 2003; Heckman, Lyons, and Todd 2000; Smith and Welch 1989], educational achievement [Jencks and Phillips 1998], out-of-wedlock childbearing [Ventura and Bachrach 2000], health (see Kington and Nickens [2001]), and criminal involvement [Reno et al. 1997]. The degree of residential segregation by race, though lower today than in 1970, remains high [Cutler, Glaeser, and Vigdor 1999; Massey 2001].

Racial differences also persist, and in some cases have become even more pronounced, on a wide range of cultural dimensions including musical tastes [Waldfogel 2003], linguistic patterns [Wolfram and Thomas 2002], and consumption choices. For instance, the cigarette brand Newport has a 75 percent market share among Black teens, but just 12 percent among White teen smokers; 65 percent of White teens smoke Marlboro compared with only 8 percent of Blacks [Johnston et al. 1999]. Seinfeld, one of the most popular sitcoms in television history among whites, never ranked in the top 50 among Blacks. Indeed, of the top ten shows with the highest viewership among Whites during the 1999–2000 television season, only one show was also among the top ten for blacks: NFL Monday Night Football (Nielsen Media Research: http://www.nielsenmedia.com/ethnicmeasure/).

Understanding whether cultural differences are a *cause* of continued economic disparity between races is a question of great social importance. Cultural differences may be a cause of Black economic struggle if Black culture interferes with the acquisition of human capital or otherwise lowers the labor market productivity of Blacks (as argued in the culture of poverty paradigm in sociology; see Hannerz [1969], Lewis [1966], Riessman [1962], and implicitly, Anderson [1990]). For instance, high-achieving Black children may be ostracized by their peers for "acting white," potentially leading to lower investment in human capital

[Fordham and Ogbu 1986; Austen-Smith and Fryer 2003]. Speaking "Ebonics" may interfere with the ability to interact with White coworkers and customers, or disrupt human capital acquisition more directly [Orr 1989]. On the other hand, the presence of a Black culture may simply be the *consequence* of past and current segregation and economic inequality, but play no role in perpetuating economic disparity. If differences in tastes do not influence human capital acquisition or labor market productivity, then there is little reason to believe that such tastes will have a causal negative economic impact on Blacks. For example, "soul food" [Counihan and Van Esterik 1997] and traditional African-American spirituals [Jackson 1944] can be traced to the social conditions endured during slavery, but are unlikely to be causes of current poverty. Eliminating cultural differences in this scenario would have no overall impact on Black welfare relative to Whites.

A primary obstacle to the study of culture has been the lack of quantitative measures. In this paper we focus on one particular aspect of Black culture—the distinctive choice of first names—as a way of measuring cultural investments.[1] Our research builds upon a growing literature by economists devoted to understanding a diverse set of social and cultural phenomena [Akerlof and Kranton 2000; Berman 2000; Fryer 2003; Glaeser, Laibson, and Sacerdote 2002; Iannaccone 1992; Lazear 1999]. In contrast to these earlier papers, however, our contribution is primarily empirical.

Using data that cover every child born in California over a period of four decades, our analysis of first names uncovers a rich set of facts. We first document the stark differences between Black and White name choices in recent years.[2] For example, more than 40 percent of the Black girls born in California in recent years received a name that not one of the roughly 100,000 White girls born in California in that year was given.[3] Even

1. Other than the audit studies of resumes discussed below, the only other economic analysis of name choices that we are aware of is Goldin and Shim [2003] which examines the issue of women retaining their maiden names at marriage. The seminal work on names outside of economics has been done in a series of papers by Stanley Lieberson and coauthors, culminating in Lieberson [2000].

2. There are multiple dimensions along which a name can be considered "black" or "white." For example, Lieberson and Mikelson [1995] study distinctive patterns of phonemes that are characteristic of Black names. In this paper we study only one dimension of the issue: the relatively frequency with which Blacks and Whites choose a given name for their children.

3. Lieberson and Mikelson [1995], using a sample of names from birth records in Illinois, find that approximately 30 percent of black baby girls born

among popular names, racial patterns are pronounced. Names such as DeShawn, Tyrone, Reginald, Shanice, Precious, Kiara, and Deja are quite popular among Blacks, but virtually unheard of for Whites.[4] The opposite is true for names like Connor, Cody, Jake, Molly, Emily, Abigail, and Caitlin. Each of those names appears in at least 2,000 cases (between 1989–2000), with less than 2 percent of the recipients Black.[5] Overall, Black choices of first names today differ substantially more from Whites than do the names chosen by native-born Hispanics and Asians.

More surprising, perhaps, is the time series pattern of Black first names. In the 1960s the differences in name choices between Blacks and Whites were relatively small, and factors that predict distinctively Black names in later years (single mothers, racially isolated neighborhoods, etc.) have much lower explanatory power in the 1960s. At that time, Blacks who lived in highly racially segregated neighborhoods adopted names that were almost indistinguishable from Blacks in more integrated neighborhoods and similar to Whites. Within a seven-year period in the early 1970s, however, a profound shift in naming conventions took place, especially among Blacks in racially isolated neighborhoods. The median Black female in a segregated area went from receiving a name that was twice as likely to be given to Blacks as Whites to a name that was more than twenty times as likely to be given to Blacks. Black male names moved in the same direction, but the shift was less pronounced. On the other hand, among a subset of Blacks encompassing about one-fourth of Blacks overall and one-half of those in predominantly White neighborhoods, name choices actually became more similar to those of Whites during this same period.

We argue that these empirical patterns are most consistent with a model in which the rise of the Black Power movement influenced Black identity. Other models we consider, such as ignorance on the part of Black parents who unwittingly stigma-

---

between 1920 and 1960 have unique names. Starting in the early 1960s, there was a remarkable increase in the prevalence of unique names, resulting in a peak in 1980 in which 60 percent of Black girls were given unique names. A similar, though less pronounced, phenomenon existed among Black boys.

4. There are 463 children named DeShawn, 458 of whom are Black. The name Tyrone is given to 502 Black boys and only 17 Whites. 310 out of 318 Shanice's are Black, as are 431 out of 454 girls named Precious, and 591 out of 626 girls named Deja.

5. The most extreme case is for the name Molly, in which only 9 of 2,248 children given the name are Black.

tize their children with such names, simple price theory models, and signaling models, all contradict the data in important ways.

The paper concludes by analyzing the relationship between distinctively Black names and life outcomes. Previous studies have found that distinctively Black names are viewed negatively by others (e.g., Busse and Seraydarian [1977]). Most persuasive are audit studies in which matched resumes, one with a distinctively Black/ethnic minority name and another with a traditionally White name, are provided to potential employers [Jowell and Prescott-Clarke 1970; Hubbick and Carter 1980; Brown and Gay 1985; Bart et al. 1997; Bertrand and Mullainathan 2003]. Such studies repeatedly have found that resumes with traditional names are substantially more likely to lead to job interviews than are identical resumes with distinctively minority-sounding names. The results suggest that giving one's child a minority name may impose important economic costs on the child. In our data, however, we find no compelling evidence of a negative relationship between Black names and a wide range of life outcomes after controlling for background characteristics. Although seemingly in conflict with prior audit studies using Black names on resumes, there are three interpretations of the data that reconcile the two sets of results: (1) Black names are used as signals of race by discriminatory employers at the resume stage, but are unimportant once an interview reveals the candidate's race, or (2) Black names provide a useful signal to employers about labor market productivity after controlling for information on the resume, or (3) names themselves have a modest causal impact on job callbacks and unemployment duration that we are unable to detect.

The remainder of the paper is structured as follows. Section II describes the data used in the analysis. Section III summarizes the basic patterns observed in the data. Section IV attempts to reconcile the stylized facts with a range of potential theories. Section V analyzes the relationship between names and life outcomes and attempts to reconcile our results with previous audit studies. Section VI concludes. A data appendix describes the details of our sample construction.

II. The Data

The data used in this paper are drawn from the Birth Statistical Master File maintained by the Office of Vital Records in the California Department of Health Services. These files provide

information drawn from birth certificates for all children born in California over the period 1961–2000—over sixteen million births. With the approval of the California Committee for the Protection of Human Subjects, personal identifiers including mother's first name, mother's maiden name, and child's full name have been added to the public use versions of the data. Details of the data set are provided in the appendix.

The information included varies by year and has generally been increasing over time. For our entire sample, we have information on the baby's first name, race, gender, date of birth, hospital of birth, and birth weight, as well as the mother's maiden name, parental ages, and inferred marital status. By 1989, information on parental education, residential zip code, and source of payment were added to the data. Also starting in 1989 we know the mother's first name and exact date of birth, two critical elements for linking information from a woman's own birth certificate to that of her children in those cases in which a woman is both born in California and later gives birth there. This allows us to look at the relationship between circumstances at a woman's birth (e.g., her own name, her mother's level of education, her mother's marital status, racial segregation, etc.) and the situation in which that same woman lives in when she gives birth decades later.[6] It also enables us to link information from all births to the same mother that took place in California, which permits the comparison of naming patterns controlling for mother-fixed effects.

We restrict our sample to non-Hispanic Blacks (referred to simply as Blacks).[7] In determining how Black a name is, our comparison group is non-Hispanic Whites. Summary statistics for the Black babies are provided in Table I.[8] We divide the sample into four sets of years: 1961–1967, 1968–1977, 1978–1988, and

6. Unfortunately, the father's first name is not included in the data, so a parallel analysis cannot be performed for males.

7. There is some ambiguity in racial and ethnic categorizations when children are born to parents of different races. We use the classification of a child's race and ethnicity on the birth certificate to assign these categories, except in the years 1999 and 2000 when the information on child's race is missing. In those years, a child is considered Black if either parent is Black. Before 1989 we do not have explicit identifiers for Hispanics. Using information from later years of data, however, we are able to effectively identify Hispanic surnames and drop from the sample any child born to a parent with a surname or maiden name that is more than 10 percent Hispanic.

8. The percentages of black babies born in a hospital or county are time varying variables measured on an annual basis using the full sample of California births. Within our four groupings of years, we provide the mean over the relevant years.

TABLE I
SUMMARY STATISTICS FOR BLACK BIRTHS, CALIFORNIA, 1961–2000

Variables	1961–1967	1968–1977	1978–1988	1989–2000
Female	0.49	0.49	0.49	0.49
	(.50)	(.50)	(.50)	(.50)
Age of mother at time of birth	24.43	23.25	24.43	25.70
	(6.12)	(5.56)	(5.47)	(6.16)
Age of father at time of birth	28.47	26.85	27.72	28.70
	(7.49)	(7.21)	(7.29)	(7.60)
Mother born in California	0.14	0.39	0.51	0.62
	(0.35)	(0.49)	(0.50)	(0.48)
Mother unmarried at time of birth	0.31	0.46	0.56	0.58
	(0.46)	(0.50)	(0.50)	(0.49)
Birth weight (in grams)	3098.87	3127.58	3181.53	3223.00
	(581.26)	(598.12)	(587.49)	(591.27)
Total number of children	3.20	2.29	2.11	2.29
	(2.36)	(1.69)	(1.34)	(1.50)
Months of prenatal care	5.87	6.77	7.08	7.28
	(1.95)	(1.68)	(1.59)	(1.59)
Teen mother at time of birth	0.23	0.29	0.20	0.18
	(0.42)	(0.45)	(0.40)	(0.38)
County hospital	0.42	0.16	0.11	0.09
	(0.49)	(0.37)	(0.32)	(0.29)
White mother	—	0.10	0.13	0.20
	—	(0.30)	(0.33)	(0.40)
White father	—	0.02	0.02	0.04
	—	(0.14)	(0.15)	(0.19)
Mother's years of education	—	—	—	12.58
	—	—	—	(1.99)
Father's years of education	—	—	—	12.60
	—	—	—	(2.55)
Privately insured	—	—	—	0.43
	—	—	—	(0.49)
Per capita income in zip code	—	—	—	13166.30
(All residents in 1989, 1989 dollars)	—	—	—	(5277.24)
Per capita income in zip code	—	—	—	11577.52
(Blacks in 1989, 1989 dollars)	—	—	—	(4190.11)
Percent of Black population in zip code	—	—	—	37.84
	—	—	—	(34.55)
Percent of Black babies in birth county	12.83	17.65	23.57	27.25
	(4.18)	(6.73)	(8.88)	(9.97)
Percent of Black babies in birth hospital	37.94	36.86	40.11	41.47
	(28.02)	(27.63)	(27.50)	(27.65)
Black Name Index	60.87	66.65	68.13	70.49
	(27.73)	(31.27)	(31.30)	(31.49)
Black Name Index—Median	60.19	70.24	73.40	82.28
Number of observations	164,648	253,735	402,120	488,959

All data are drawn from California birth certificate records, except for the information on zip codes which combines birth certificate information with data from the 1990 Census. The sample in columns 1–4 represent all Blacks born in California between 1961–1967, 1967–1977, 1978–1988, and 1989–2000, respectively. The numbers in parentheses are standard deviations. See the data appendix for further details of the construction of the samples and variables.

1989–2000. The cutoffs for these groups have been chosen with three goals in mind: (1) roughly equalizing the time periods in each group, (2) matching the cutoffs to breaks in trend in the aggregate data, and (3) grouping years in which similar sets of covariates are available. Excluded from the data set are a small percentage of observations missing information on names. When other variables are missing, we opt to leave the observation in the analysis, including an indicator variable for a missing value.

The bottom row of Table I presents our summary measure of how distinctively Black first names are. The measure we choose, which we term our "Black name index" (BNI) is

$$(1) \quad BNI_{name,t} = \frac{\Pr(name|Black,t)}{\Pr(name|Black,t) + \Pr(name|White,t)} * 100,$$

where *name* reflects a particular first name and *t* reflects the year of birth. The index ranges from 0 to 100. If all children who receive a particular name are White, then BNI takes on a value of 0. If only Black children receive a name, BNI is equal to 100. If Whites and Blacks are equally likely to choose a name, BNI equals 50. If Blacks are four times as likely to select a name, then BNI takes on a value of 80 (e.g., .4/[.4 + .1] * 100 = 80).[9] A BNI of 90 implies Blacks choose a name nine times more often. This measure is invariant to the fraction of the population that a particular minority group comprises, and to the overall popularity of a name. Names that are pronounced the same but spelled differently are treated as separate names. The bottom row of Table I shows that the mean BNI for Blacks rises from 60.9 in the early period (implying that the mean Black name is given to Blacks about 50 percent more often than it is given to Whites) to 71.0 in the last period (implying that these names are given to Blacks about two and one-half times more frequently than Whites).

Figure I(A) more clearly demonstrates the dramatic differences between Black and White name choices. The figure presents a smoothed plot of the probability distribution function of Black and White names. The horizontal axis reflects how Black an individual's name is. The vertical axis measures the density of names chosen by race. More than 40 percent of Whites are given names that are at least four times as likely to be given to Whites

9. In computing BNI for a particular child, we exclude that child from the calculation. When a name is unique, meaning that only one child receives that name in a particular year, we assign the name a value of 0 if the person getting the name is White and 100 if the baby is Black. We explore unique names in greater detail later in the paper.

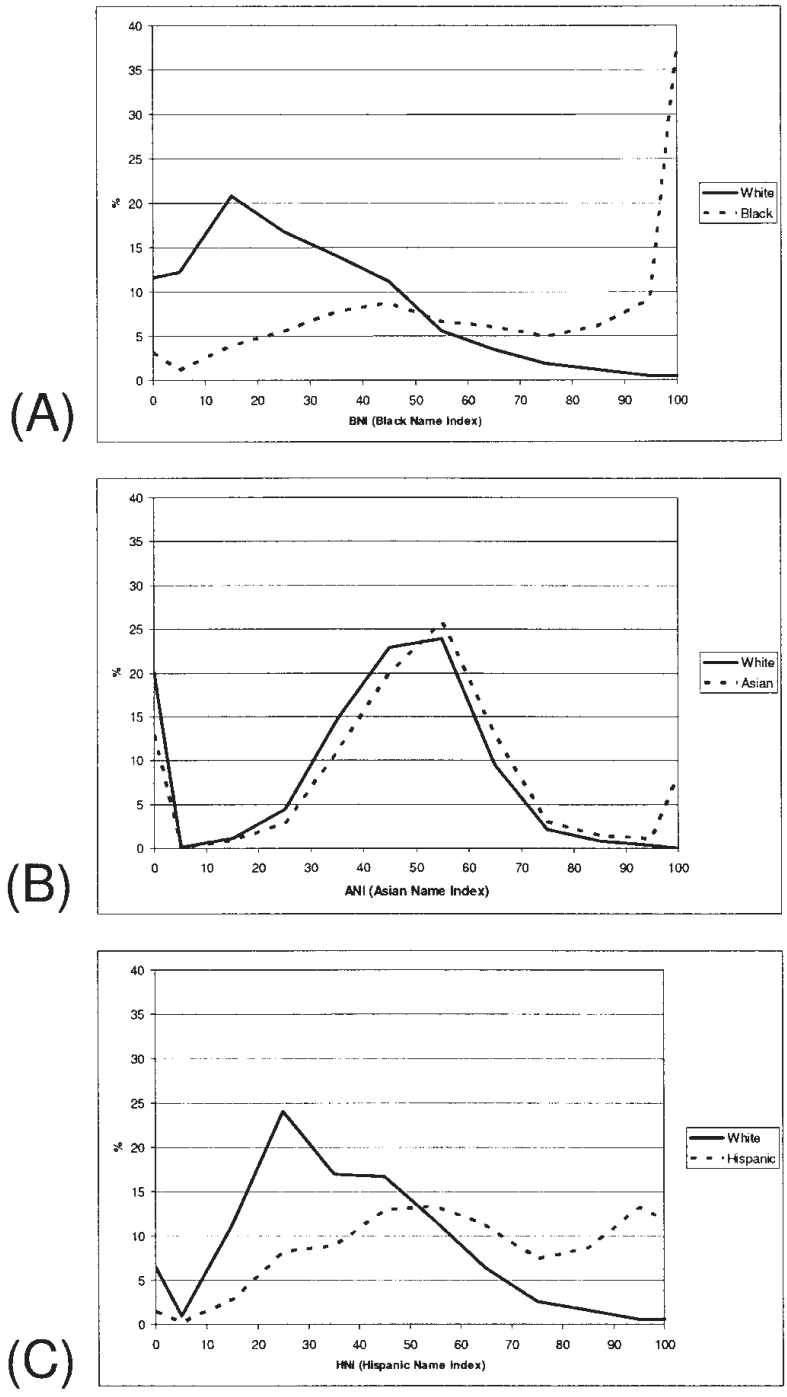


FIGURE I
Black Name Index (A), Asian Name Index (B), and Hispanic Name Index (C) Distributions, 1989–2000

(between 0–20). The fraction of Whites steadily shrinks as one moves from left to right in the figure. More than half of all Blacks have names that are at least four times as likely to be given to Blacks (between 81–100). For both races there is very little weight in the middle of the distribution (41–60), implying that there are relatively few individuals carrying names that are similarly likely for Blacks and Whites.

One might suspect that the sharp differences across races in Figure I(A) may in part be an artifact of how we construct our measure of Black names using the observed empirical distribution. In other words, we might miscategorize a name as being distinctively Black or White simply because for many names we observe only a few individuals with that name. Limiting the sample to names that appear at least twenty times in the data, however, does little to change the picture. Figure I(B), which is identical to Figure I(A) except that it compares the naming patterns of Whites with that of American-born Asians, further demonstrates that the result for Blacks is not an artifact of our measure. With the exception of a small fraction (approximately 10 percent) of the Asian population adopting names that are rare among Whites, name choices of American-born Asians strongly parallel White name choices. A comparison of native-born Hispanics and Whites in Figure I(C) shows differences in naming patterns among these two groups, although there is still substantially more overlap than for Blacks and Whites.[10]

An important racial difference in naming patterns is the greater usage of unique or nearly unique names in the Black community (see also Lieberson and Michelson [1995]). Figures II(A) and II(B) report, by race and gender, the number of children born in California in that same year (regardless of race) with that child's name. Remarkably, nearly 30 percent of Black girls receive a name that is unique among the hundreds of thousands of children born annually in California. Among Whites, that fraction is only 50 percent. Similarly, the fraction of unique names among Black boys is six times higher than for White boys, although only about half the rate of Black girls. The median Black child shares

10. We have also compared the names chosen by Whites with different levels of education. There are systematic differences in name choices (larger, in fact, than between Asians and Whites overall), but these differences are much smaller than either the Black-White or Hispanic-White gap.

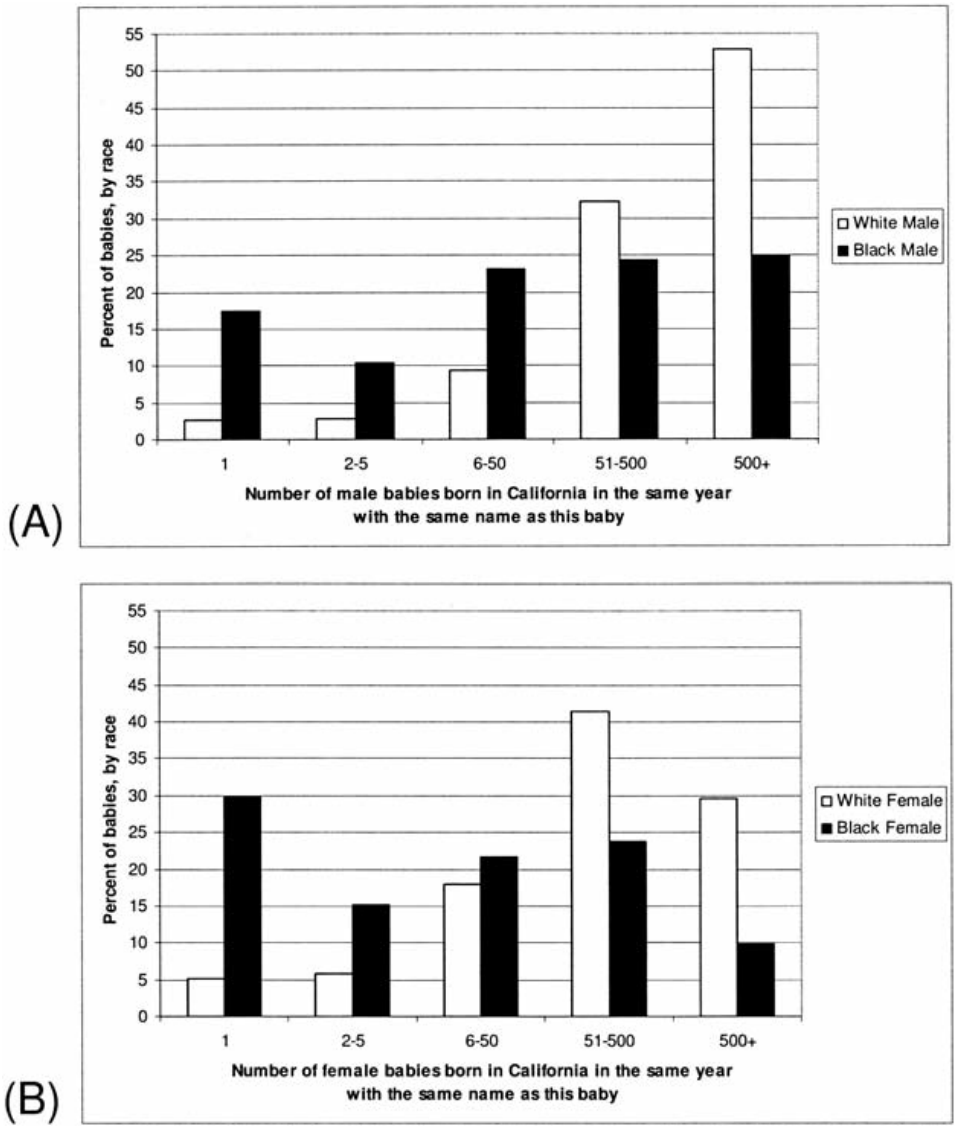


FIGURE II
Distribution of Male (A) and Female (B) Babies by How Many Share a Name, 1989–2000
(among children of all races born in California in a year)

his or her name with 23 other children; the number is almost fifteen times greater for Whites (351).[11]

Perhaps the most interesting findings in the data are the

11. The differences between Blacks and Whites in Figure II are not primarily due to the fact that there are many more Whites than Blacks born in California each year. If we randomly select a subset of Whites equal in size to the number of Blacks born each year, a similar pattern of results persists.

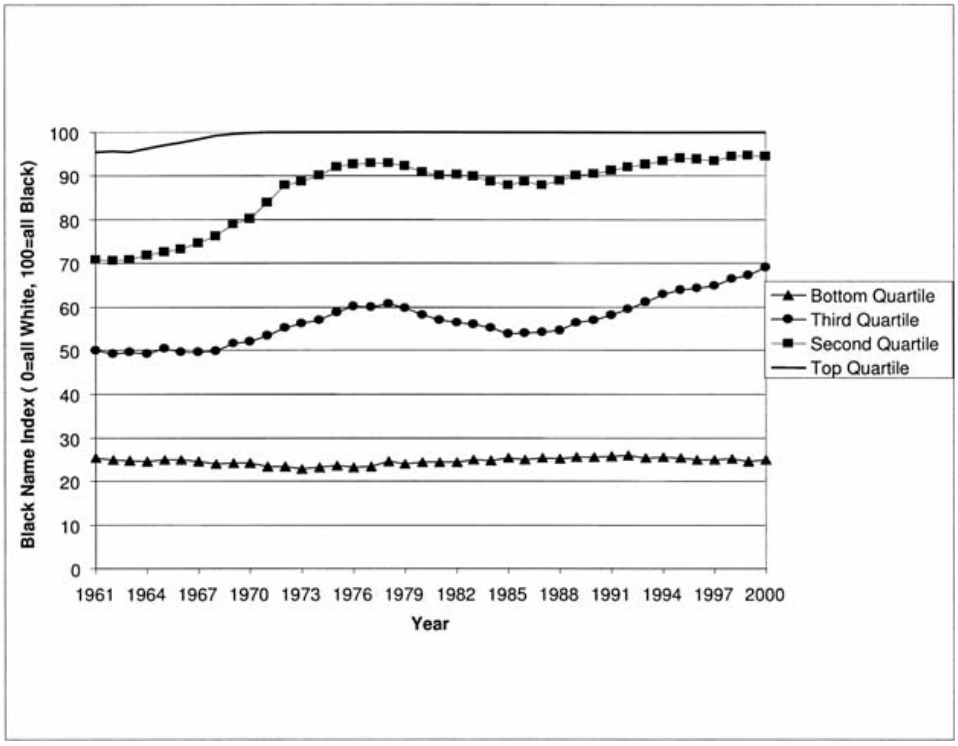


FIGURE III
Black Naming Patterns, 1961–2000 by Quartile of Black Name Index

changes in the distribution of name choices over time. For children born in each year between 1961 and 2000, we compute our Black name index and then rank order the Blacks in our sample according to how Black their name is. Figure III presents the mean BNI by year for each of the four quartiles of the distribution. The top quartile is very close to 100 throughout the entire time period (i.e., almost one-quarter of Blacks had names virtually never given to Whites throughout the sample) and thus exhibits little time-series variation. For the other three quartiles, Black naming patterns were largely stable throughout most of the 1960s. Beginning in the late 1960s, the second quartile from the top experiences a sharp rise in how Black the name choices are. Between 1968 and 1977, the mean BNI within this quartile goes from roughly 75 (meaning the name was three times as likely to be given to a Black baby as a White baby) to almost 95 (fifteen times more likely to be given to a Black baby). The third quartile also rises over that time period, but not as sharply, and also steadily increases over the period 1985–2000. The bottom

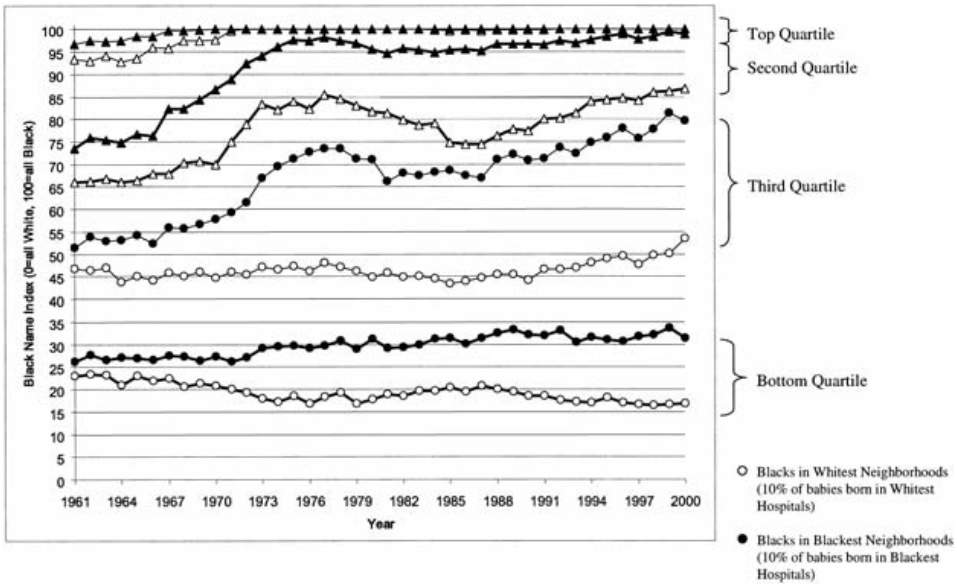


FIGURE IV
Changes in Black Naming Patterns by Racial Composition of Neighborhood and by Quartile of Black Name Index

quartile, in contrast, remains almost unchanged throughout the sample period.[12]

Figure IV is identical to Figure III, except that it disaggregates the data by the degree of racial segregation in the hospital in which a Black baby is born. We sort by racial segregation of the hospital and then by quartile of BNI. We show only the top and bottom deciles (i.e., the 10 percent of Blacks giving birth in hospitals with the greatest percent Black and the lowest percent Black, respectively) to highlight the extremes of the distribution. The most important observation emerging from Figure IV is the widening gap between the names given to Black babies born in predominantly Black and predominantly White hospitals over time. At the beginning of the sample, name choices differ little by the degree of racial segregation. For example, in the bottom quartile, Blacks born in racially segregated hospitals have a mean BNI of 26 versus

12. The divergence in naming patterns in the late 1960s and early 1970s is unique to Blacks. Among Hispanics and Asians, name choices relative to Whites change little between 1961 and 1978. This suggests that a shift in Black name choices drives the patterns in Figure III. In the 1990s, however, Hispanic names diverge somewhat from those of Whites, suggesting that changes in White naming patterns may be a factor in that time period.

23 for Blacks born in predominantly White hospitals. By 1978 these gaps have widened substantially, particularly in the third quartile where Blacks in Black hospitals have shifted more than twenty points relative to the beginning of the sample, but names of Blacks in White hospitals are essentially unchanged. Interestingly, among the Black babies given the least distinctively Black names (the bottom quartile), those born in White hospitals actually see a discernible *decrease* in how Black their names are, in contrast to the rest of the distribution.[13]

The sharp rise in the prevalence of distinctively Black names is driven in part by the increasing use of unique names among Blacks. Figures V(A) and V(B) present time-series data on the percentage of babies with unique first names by gender and racial mix of the birth hospital, where a unique name is defined as a name that no other baby born in California that year shares. The fraction of Black baby girls receiving unique names rose for the first fifteen years of the sample. Initially, rates of unique naming were similar among Blacks in predominantly Black and predominantly White neighborhoods (around 10 percent); that gap grows over time. Nonetheless, even for Blacks in White neighborhoods, almost one-quarter of baby girls received unique names. Among Whites, the rates are around 5 percent, although this (as well as the Black numbers) is likely to be an upper bound due to typographical errors being counted as unique names.[14] Figure V(B) documents a parallel, though less pronounced, phenomenon for Black boys. In 1961 the percentage of unique names ranged from 5–6 percent, irrespective of hospital. By 2000, one-fifth of Black boys born in predominantly Black hospitals receive unique names.

Unique names, while an important part of the explanation for the divergence in Black and White naming patterns, are not the entire story. Figure VI presents time-series evidence on the fraction of Blacks who have nonunique names that are above 80 on the BNI index, as well as the fraction of Blacks with unique names. The share of babies given distinctively Black, but nonunique names rises from 15 percent to 22 percent from the be-

13. The choice of names by Whites over time does not exhibit this pattern of bifurcation that is present in Black naming in Figures IV and V. The BNI for Whites is almost uncorrelated with the racial segregation of the hospitals in which they give birth throughout the entire time period.

14. There are two difficulties that arise in trying to purge the data of typographical errors. First, the scale of the data is enormous. Second, we have no way of knowing whether a particular name represents a typo (e.g., "Brian" mistyped as "Brain") or linguistic innovation.

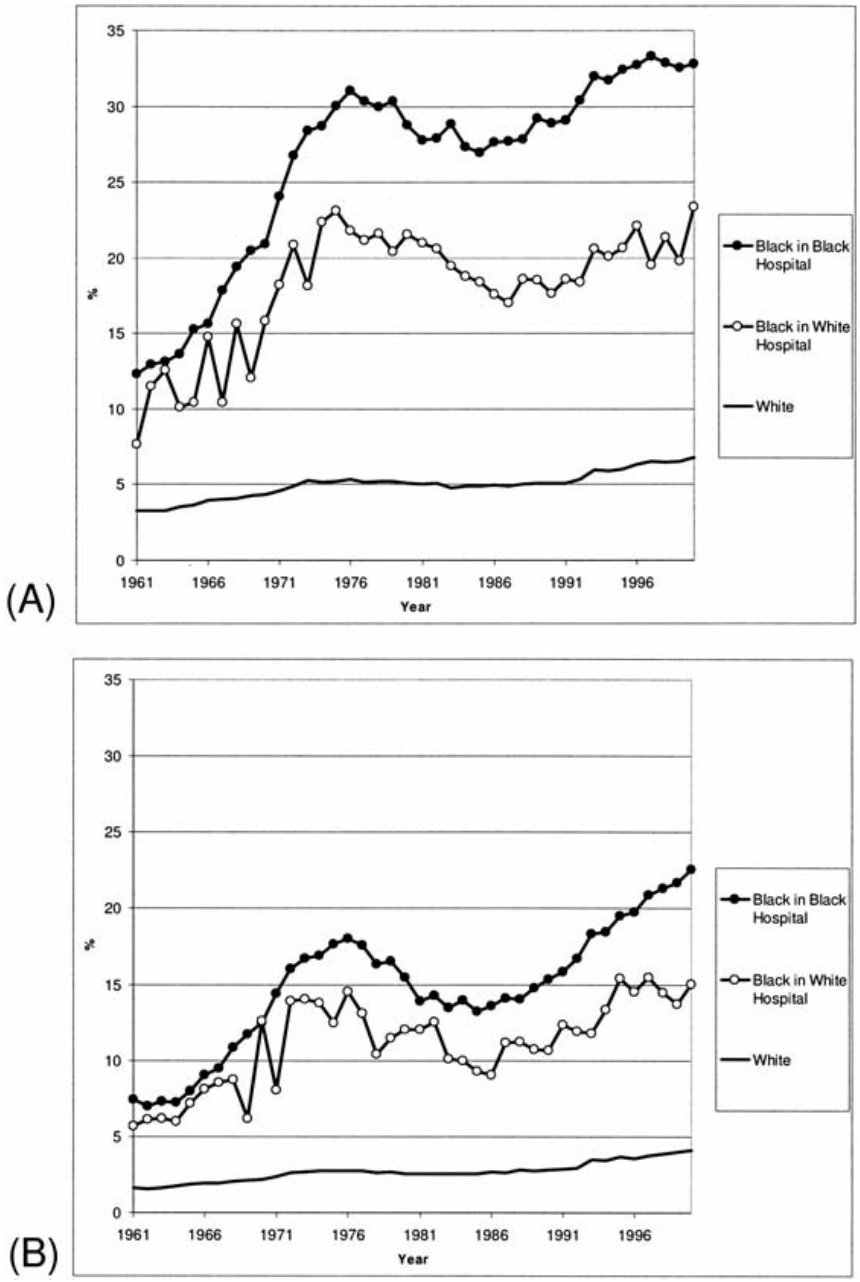


FIGURE V
Percentage of Unique Girl (A) and Boy (B) Names by Racial Mix of Hospital

ginning of the sample to the peak in the mid-1970s. Unique names rise from 7 percent to 18 percent over that same period. After a dip in the 1980s, unique names gain in prominence in the 1990s, whereas nonunique distinctively Black names do not rise in popularity in the later years.

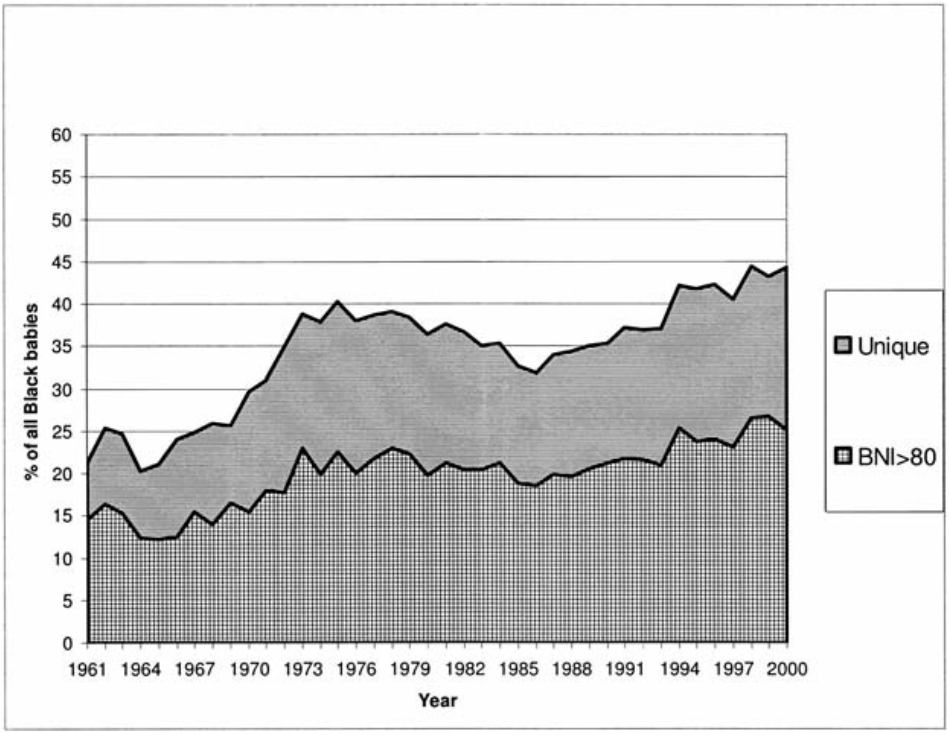


FIGURE VI
Percent of Black Babies Given Unique Names and Nonunique Distinctively Black Names, 1961–2000

Table II examines the relationship between parental characteristics, birth circumstances, and the names given to children at four different points in time. We estimate equations of the form:

$$BNI_{iht} = X_{iht}\beta + Z_{ht}\delta + Y_{ct}\theta + \varepsilon_{iht}, \quad (2)$$

where i, h, and t index individuals, hospitals, and time, respectively; X_{iht} represents an array of individual level background variables, Z_{ht} are hospital-level (potentially time varying) controls, and Y_{ct} denotes county-level (possibly time varying) controls. The standard errors are clustered by hospital-year. The first four columns of the table correspond to different sample periods. Column 1 reflects Black children born between 1961 and 1967, prior to the sharp changes in Black naming behavior. Column 2 has births occurring between the years 1968 and 1977, the years in which the transition occurred. Columns 3 and 4 capture the periods 1978–1988 and 1989–2000, respectively. We separate these two time periods because of the availability of a more complete set of covariates after 1989. Also, after 1989 we have the

necessary information to link multiple births to the same women, allowing for the inclusion of mother-fixed effects in column 6. All of the controls included in the regression are taken from a child's birth certificate and thus are determined at the time of birth. For the entire period, covariates available include parents' marital status, age, and race, mother's place of birth, her total number of children, months of prenatal care, the baby's birth weight, the percent Black at the birth hospital, the percent Black in the county, and whether the child is born in a county hospital. In the later years of the sample, the set of included controls is much richer: income and percent Black at the zip code level, highest grade completed by mother and father, and expected source of payment for the delivery.

Columns 1–4 hold the set of covariates constant and compare the relative importance of these variables in predicting BNI over the four periods. In almost all cases, variables associated with low socioeconomic status are also associated with Blacker names. Moreover, the link between low socioeconomic status and Black names becomes much stronger over time. The coefficients on mother's age, birth weight, and single mother are less than one-half as large in the first period (column 1) as they are in the last period (column 4). Father's age has no impact in the early period, but does have a negative coefficient in the later periods. Note also that the R^2 in the regressions increases steadily over time, meaning that these characteristics explain a growing fraction of the variation in names. A final implication of the table is that names are becoming more distinctively Black over the course of the sample. For instance, a woman with the mean sample characteristics in the 1960s would be predicted to choose a name with a BNI of 60.9 in the 1961–1967 period. A woman with those same characteristics giving birth in the 1990s will on average choose a name with a BNI of 71.4. Likewise, for a woman with the average traits of mothers giving birth in the 1990s, the predicted BNI of the name is 70.5 for a baby born in the 1990s compared with 61.8 if that baby had been born in the 1960s.[15]

The specification estimated in columns 5 and 6 are similar to those in the previous columns of the table, except the richer set of

15. The results in Table II are similar if the sample is restricted to names given to twenty or more babies in a given year. Patterns are also similar if we use as the dependent variable an indicator for whether or not a child has a unique name. Although not shown in tabular form, we also find that many of the same factors that predict the choice of names among Blacks also predict the choice of names among Whites, although for Whites the magnitude of the impacts is substantially smaller in magnitude.

TABLE II
DETERMINANTS OF NAME CHOICES AMONG BLACKS

Variable	1961–1967	1968–1977	1978–1988	1989–2000	1989–2000 Full covs	1989–2000 Mother FEs
Female	1.29	3.31	4.84	3.06	3.17	4.04
	(0.14)	(0.14)	(0.11)	(0.10)	(0.10)	(0.14)
Percent of Black babies in birth hospital	0.045	0.060	0.081	0.091	0.027	0.018
	(0.004)	(0.005)	(0.003)	(0.004)	(0.004)	(0.005)
Percent of Black babies in county	0.064	0.142	0.035	0.012	0.008	−0.004
	(0.020)	(0.016)	(0.010)	(0.010)	(0.010)	(0.016)
Mother born in California	−2.16	0.46	−0.06	0.46	0.11	—
	(0.35)	(0.16)	(0.11)	(0.11)	(0.10)	—
Mother's age at time of birth	−0.35	−0.54	−0.63	−0.74	−0.50	0.15
	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.05)
Father's age at time of birth	0.00	−0.17	−0.13	−0.19	−0.16	−0.06
	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
Months of prenatal care	−0.07	0.27	−0.04	−0.17	−0.04	0.00
	(0.04)	(0.04)	(0.03)	(0.03)	(0.03)	(0.05)
County hospital	1.16	−0.05	0.68	0.08	−0.90	−0.60
	(0.20)	(0.29)	(0.19)	(0.30)	(0.29)	(0.35)
Birth weight in kilograms	−0.42	−1.12	−1.91	−1.78	−0.59	0.02
	(0.12)	(0.11)	(0.09)	(0.08)	(0.08)	(0.16)
Total number of children	0.22	0.73	1.32	1.84	0.85	−0.06
	(0.04)	(0.05)	(0.05)	(0.04)	(0.04)	(0.11)
Single mother	2.10	4.21	5.21	4.73	2.95	0.20
	(0.28)	(0.15)	(0.13)	(0.12)	(0.11)	(0.21)
Percentage of Black babies in zip code	—	—	—	—	−0.0029	−0.0103
	—	—	—	—	(0.0024)	(0.0045)

TABLE II
(CONTINUED)

Variable	1961–1967	1968–1977	1978–1988	1989–2000	1989–2000 Full covs	1989–2000 Mother FEs
Per capita income (all residents, 1989)	—	—	—	—	−0.0622	−0.0116
(× 1000)	—	—	—	—	(0.0163)	(0.0340)
Per capita income (Black residents, 1989)	—	—	—	—	−0.2294	−0.0830
(× 1000)	—	—	—	—	(0.0187)	(0.0389)
Mother's years of education	—	—	—	—	−0.850	0.060
	—	—	—	—	(0.030)	(0.073)
Father's years of education	—	—	—	—	−0.295	−0.029
	—	—	—	—	(0.023)	(0.043)
Privately insured	—	—	—	—	−2.98	−0.55
	—	—	—	—	(0.15)	(0.22)
White mother	—	—	—	—	−12.22	—
	—	—	—	—	(0.13)	—
White father	—	—	—	—	−17.94	−6.28
	—	—	—	—	(0.26)	(0.81)
Constant	66.61	76.11	81.61	88.18	104.61	—
	(0.78)	(0.59)	(0.51)	(0.50)	(0.62)	—
R^2	0.0117	0.0354	0.0464	0.0592	0.0951	0.7764
Birth sample used	1961–1967	1968–1977	1978–1988	1989–2000	1989–2000	1989–2000
Include mother-fixed effects?	No	No	No	No	No	Yes
Number of observations	164648	253703	402120	488959	488959	488485

The dependent variable in all columns is the Black Name Index of the name given to the child at birth. The Black Name Index ranges from 0 for names that are only given to Whites to 100 for names that are only given to Blacks. All variables included on the right-hand side of the regression are known at the time of a baby's birth. Included in the regression, but not reported in the table are indicator variables for missing values. Standard errors are clustered to take into account correlation across people born in the same hospital in the same year.

covariates is used. Blacker names are associated with lower-income zip codes, lower levels of parental education, not having private insurance, and having a mother who herself has a Blacker name.

The last column of Table II adds mother-fixed effects to the specification. Thus, the identification comes from changes in a mother's situation across different births. To the extent that mothers anticipate the changes that occur in their circumstances between births and factor these expectations into the name given to their first child, the fixed-effects coefficients will be biased toward zero. The estimates in column 6 are substantially smaller in magnitude than in the earlier columns, with the exception of variables reflecting how Black the neighborhood is and the coefficient on female. The coefficient on mother's age flips sign, implying that women who have babies early in life tend to choose Black names relative to other women, but that a given woman picks slightly Blacker names for her children as she grows older.[16] In addition, the inclusion of mother-fixed effects raises the R^2 from .095 to .776—there is a great degree of continuity in the names mothers choose for their children, implying that person-specific tastes are quite important determinants of name choice. The results with mother-fixed effects suggest either that temporary changes in circumstances have a relatively small impact on name choices, or that our measures of current circumstances are noisy.

III. Understanding the Patterns in the Data through the Lens of Economic Theory

A number of stylized facts emerge from the analysis of the preceding sections. Blacks, much more than other minorities, choose distinctive names for their children. The distinctiveness of Black names has risen greatly over time, most notably in the late 1960s and early 1970s. These shifts in naming patterns have not, however, been uniform. In particular, among the quarter of the Black population choosing the names most common among Whites, the opposite pattern is evident. Further, Blacks living in highly segregated Black communities today are much more likely to have distinctively Black names than those in integrated communities, whereas this was not the case in the early 1960s.

16. This result holds up when we include a full set of year dummies, implying that the increases in BNI are age effects of the mothers and not time effects.

Finally, until the late 1970s the choice of Black names was only weakly associated with socioeconomic status; in the 1980s and 1990s distinctively Black names have come to be increasingly associated with mothers who are young, poor, unmarried, and have low education.

In this section we consider the extent to which existing theories can successfully account for this disparate set of facts.

III.A. The Ignorance Model

One explanation for the prevalence of distinctively Black names is ignorance on the part of Black parents, who fail to appreciate the costs they are imposing on their children through such choices. Audit study results, for instance, suggest that Black names may be punished in the labor market.

This theory fails to adequately explain the sharp increase in distinctively Black names in the late 1960s and early 1970s. There is little reason to believe that Black parents became systematically less informed about the consequences of distinctively Black names at that time. This theory is also at odds with the fact that those adopting Black names in the early 1970s were, for the most part, representative of the Black community, not a small subset of parents likely to be particularly misinformed.[17]

III.B. Price Theory Model

Consider the following skeletal outline of a price theory model of names, which we derive formally in an earlier version of this paper [Fryer and Levitt 2002]. Parents give names to their children at birth to maximize the child's expected utility. Individuals are born into neighborhoods that differ in racial composition. People live and work in the same neighborhoods. Moving between neighborhoods is costly. The returns to ability are assumed to be higher in predominantly White neighborhoods. White names pro-

17. On the other hand, one cannot a priori rule out that this explanation has potential relevance for explaining at least some of the patterns of the last two decades. If it became apparent in the 1980s that there were costs to having a Black name, one might expect that parents who were likely to be best informed about these costs (e.g., older working parents living in less racially isolated neighborhoods) would choose such names less frequently, whereas learning on this dimension might be more gradual for teenage mothers in the inner city. In absolute terms, this explanation fails, because the adoption of Black names rose throughout the 1990s. At least in relative terms, however, there was a shift in the 1980s and 1990s toward an increasing concentration of distinctively Black names among parents least likely to be well informed about the potential stigma of such names in broader society. In light of results presented later in the paper that call into question the costs of adopting distinctively Black names, however, we are quite skeptical of this theory as an explanation for the observed phenomena.

vide benefits in interactions with Whites; Black names are beneficial when interacting with Blacks. We also assume that the value of a White name is increasing in ability, motivated by the audit study literature which suggests that the primary cost of a distinctively Black name is via the labor market. A child's ability is not known with certainty at birth, although the distribution from which the child's ability will be drawn is known.

In this model, parents will opt for Whiter names when (i) their children are more likely to be high ability, (ii) the cost of moving to predominantly White neighborhoods falls, (iii) returns to ability rise in the labor market, (iv) the relative cost of having a White name when interacting with Blacks falls, and (v) the benefit of having a White name when interacting with Whites rises.

The predictions of this model face mixed success in terms of the patterns observed in the data. Consistent with the theory is the fact that those in racially isolated communities are especially likely to adopt distinctively Black names in recent decades, and that such names are most common among groups likely to face the greatest barriers to participating in traditional labor markets. This theory, however, does quite poorly in explaining the sharp rise in distinctively Black names in the late 1960s and early 1970s—a period immediately following the passage of national Fair Housing laws in 1968, falling social barriers to integration, and increased economic opportunity for Blacks. Empirically, urban racial segregation which had been rising began to fall around this time and has steadily declined for three decades [Cutler, Glaeser, and Vigdor 1999].[18] In light of these apparent reductions in the costs of switching between neighborhoods, one would have expected a shift away from distinctively Black names at precisely the point where such names were becoming most prevalent.

Additionally, it is not clear that the price theory model provides an adequate explanation as to why Black names are much more distinctive than Asian or Hispanic names, when presumably many of the same trade-offs might also exist among those minority groups.

18. Although the level of residential racial segregation has declined at a relatively slow rate, to the extent that these persistent declines were anticipated and parents are concerned with the impact a name will have on their children as adults in the labor market, one would nonetheless expect to see an immediate, abrupt shift in naming patterns.

III.C. A Signaling Model

A third model to consider is a simple signaling model in which distinctively Black names serve as a signaling device, but are otherwise not productive, along the lines of prior research by Iannaccone [1992], Berman [2000], and Fryer [2003].[19] Imagine a predominantly Black neighborhood populated by Black individuals of one or two types: those who have a strong affinity for the Black community (the "black" type) and those who do not (the "white" type). An individual's type is fixed at birth and cannot be changed. Each individual knows her own type, but type is not observable to others. An individual's utility is determined by his type and the neighborhood's perception about his type. Regardless of one's actual type, social interactions in the Black community yield higher utility if others believe that one has an affinity for the Black community.[20] Thus, all else being equal, both types prefer to be perceived as being a "black" type.

One way in which an individual signals his or her type is by the names that they choose for their children. In the model, each parent has one child and bestows a name on that child. The (suitably anthropomorphized) community observes the name that each parent gives his or her child, and based on that signal, draws unbiased inferences about the parent's type. As in the price theory model, we assume that White names provide labor market benefits. The total cost and marginal cost of giving a child a Blacker name is lower for a parent with the "black" type.

Either separating or pooling equilibria may arise in such models. In any separating equilibrium, parents whose type is "black" are willing to give otherwise costly names to their children simply because it allows them to distinguish themselves from parents who identify as "white," and derive more utility from social interactions. As a result, peers come to regard Black names as a signal of community loyalty. In any pooling equilibrium, the payoff to social interactions is determined by the underlying population distribution of types. Pooling is likely to occur, in our model, when the marginal rates of substitution for both types of parents are similar.

The predictions of the signaling model fit some of the basic

19. For a formal treatment of this model, see the earlier version of this paper [Fryer and Levitt 2002].

20. For example, those with an affinity to the Black community may be more likely to contribute to local public goods than those with "white" identities, inducing better treatment from their neighbors. For instance, Anderson [1999] talks about the importance of having others "watch your back."

patterns observed in the data. Black name choices in the 1960s, for instance, look roughly consistent with a pooling equilibrium in which one's name is not a strong signal. Any mechanism that amplified the differences in the cost of signaling between the two types could cause a bifurcation in the distribution of Black names such as occurred in the early 1970s, with "black" types moving to distinctively Black names and "white" types shifting toward Whiter names (if they previously had been in a pooling equilibrium) or continuing to choose traditional white names (if they formerly were in a separating equilibrium). Such a change could have come as a result of the increased opportunities for integration due to the civil rights movement [Fryer 2003]. Those Blacks with the most to gain from integration would opt for Whiter names. Somewhat counterintuitively, for those Blacks who continue to signal a "black" type, to sustain the new equilibrium requires an even costlier (i.e., Blacker) name choice since the outside option is now more attractive.

The signaling model appears to fall short in explaining four dimensions of the data. First, in the 1980s and 1990s, individuals who are most likely to adopt distinctively Black names (young, unmarried women in predominantly Black neighborhoods) are those with the least need to signal affinity with the Black community or their commitment to remaining in the neighborhood. Second, the signaling model has a hard time explaining why Black names became more prevalent among many different types of Blacks, including those who live in mostly White communities. Third, in order for the Black names signal to be credible, it must impose costs on those who carry the names. In Section IV of this paper, however, we are unable to detect evidence that Black names are costly. Finally, the signaling model provides little rationale as to why Blacks might engage much more extensively in signaling than do either Asians or Hispanics.

III.D. An Identity Model

The primary shortcoming of the signaling model in explaining the data is that, for a large segment of the Black community, distinctively Black names appear to be viewed as a benefit rather than a cost. In the signaling model, an investment must be costly to provide a credible signal. Using a similar framework, but allowing Black names to be a benefit for those with the "black" type and a cost for those with the "white" type converts the signaling model into a simple identity model.

In particular, we have in mind the framework of Akerlof and

Kranton [2000]. In their language, identities are accompanied by certain "prescriptions" that define appropriate behaviors for a person of that type. When an individual takes actions in line with these prescriptions (e.g., a "black" type choosing a distinctively Black name, or a "white" type choosing a White name), there is a utility benefit.

To justify the patterns in the time series circa 1970 through this model, there must be a shock to the identity prescriptions (i.e., what it means to be Black) around this time period. The rise of Black Power appears to be precisely that sort of shock. The underlying philosophy of the Black Power movement was to encourage Blacks to accentuate and affirm Black culture and fight the claims of Black inferiority [Van Deburg 1992]. Within the Black community, there were widespread changes in hair styles and the rising popularity of afro-centric clothing between 1968–1975 [Van Deburg 1992; Woodward 1974]. The adoption of distinctively Black names would be completely consistent with these other cultural phenomena. The identity model may also help to explain why naming patterns among Blacks are quite distinctive from Whites, but Asians name their children in much the same manner as Whites. For instance, if Asian "prescriptions" stress financial success and assimilation, Asian names would be expected to mirror those of Whites.

Another fact consistent with a Black Power explanation is that the concentration of Blacks by county is a much stronger predictor of Black names in the 1968–1977 period than the rest of the sample. In column 2 of Table II, the coefficient on that variable is .142 (standard error of .016) for 1968–1977, but never larger than .064 in any other period. The county-level trends also exhibit a basic consistency with a Black Power story. Alameda (in which Oakland is located), Los Angeles, and San Francisco counties were the centers of the Black Power movement. Black names increased earlier and to a greater extent in these three counties than the rest of the state.

The identity model may also apply to the increases in BNI in the 1990s, as there is suggestive evidence that there was a secondary Black cultural movement during this time. Indeed, this is precisely the period in which Blacks (headed by Jesse Jackson and Detroit Mayor Coleman Young) demanded to be referred to as African-American. This change was meant to tie Blacks more closely to their African cultural heritage. Enrollment in historically Black colleges and universities, which had been declining for a decade, rose over 20 percent between 1986 and 1994 [Hoffman

1996]. There was an emergence in the entertainment industry of politically motivated music (*Public Enemy,* for example) and film (i.e., *Do the Right Thing,* Spike Lee), which highlighted the abhorrence in black communities of the status quo. Recall, a similar spark ignited the Black Power movement, as some Blacks were frustrated with the niceties of the civil rights generation. In a related paper, McAdams, Fryer, and Levitt [2003] provide a more thorough investigation of the black power movement between 1968–1975 and the Black cultural movement in the 1990s; as explanations of the time-series changes in BNI.[21]

In summary, although the evidence in favor of the identity model is far from overwhelming, it is the only theory examined which does not yield predictions that are directly at odds with the observed patterns in the data.

IV. The Relationship between Names and Adult Outcomes

In light of audit studies documenting the use of names as a screening device by employers, one might expect that having a distinctively Black name should be associated with worse economic outcomes, holding all else equal. In this section we link information from a woman's own birth certificate to her adult circumstances as reflected in the information on the birth certificates of her children at the time she gives birth.

In order to make this linkage, a woman must be born in California and later give birth in California. We focus our analysis on women born in 1973 and 1974—the earliest years for which we have the necessary information to make reliable links. Of the women still residing in California who have given birth, we match at a high rate (over 90 percent).[22] The subset of women that we successfully link is not representative of all women born in 1973–1974. In particular, women who defer childbearing to their late twenties are excluded from our sample. The women we are able to link are themselves born to slightly younger, unmarried mothers.

21. Lieberson and Michelson [1995] make a similar point as they relate the rise in the prevalence of unique names to the rise of the Black Panther Party.

22. According to the 1990 Census, roughly 20 percent of Black women born in California live outside the state during their childbearing years. Also based on 1990 Census data, 42 percent of Black women born in California have no children by age 27. If leaving California is independent of the decision to have a child by age 27, then 46.4 percent of Black women born in California should give birth in California in our sample. [1–.2] $*$ .58. For details of how this linkage is performed, see the data appendix. A smaller subset of White women are successfully linked because more Whites move out of state and more White women defer childbearing until later in life.

The differences between the whole sample and the linked subset are consistent with higher fertility rates and lower rates of cross-state mobility among these groups. Importantly, however, after controlling for socioeconomic characteristics, there is no systematic difference between the first names of the women who are or are not successfully linked. If we regress an indicator variable for a successful link (equal to one if a link occurs and zero otherwise) on the woman's BNI and background characteristics at the time of the woman's birth, the coefficient on the woman's BNI is both substantively and statistically insignificant. Nonetheless, it is important to emphasize that our sample for testing the relationship between names and life outcomes is limited to females born in California who later give birth there before the age of 27. Further, our outcome measures are coarse. We do not observe individual wages or family economic circumstances, but rather, the median income of their zip code, years of education, etc, which are highly correlated with the relevant outcomes.

Our approach to testing for a relationship between names and life outcomes involves predicting adult life outcomes as a function of everything known about a woman and her parents at the time of her own birth, including her name:

$$y_i^{adult} = \beta BNI_i + \gamma X_i^{birth}\,\Gamma + \lambda_h^{birth} + \varepsilon_i, \tag{3}$$

where i indexes women, h represents hospitals, and the superscripts $adult$ and $birth$ correspond to the time at which the variable is measured. In some specifications we restrict the sample only to Blacks; in other cases we include both Blacks and Whites including an indicator for race and interactions between BNI and race. We analyze a wide selection of outcomes as dependent variables. All of the covariates included in the earlier analysis of the cohorts born in the 1970s are also in this specification. We limit the sample to the last birth that we observe for a particular woman in order to most closely approximate the long-run outcomes of the women, although the results are not sensitive to this restriction.

The clear weakness of this empirical approach is that if unobserved characteristics of the woman are correlated both with life outcomes and her name, our estimates will be biased. Given that Black names are associated with lower socioeconomic status on observable measures, one would expect that Black names are also likely to be positively correlated with omitted variables that

predict worse life outcomes, leading our estimates to exaggerate the true relationship between a woman's name and her life outcomes, although one can also construct scenarios in which the opposite bias could arise.[23]

The results of estimating equation (3) restricting the sample to Black women are presented for a range of outcome variable in the top two rows of Table III. The top row does not include any controls; the second row includes the full set of controls. In both cases, only the coefficient on the woman's own Black Name Index is presented in the table. In the absence of controls for background characteristics, Blacker names are uniformly associated with worse adult outcomes. Given the correlation between Black names, growing up in segregated neighborhoods, and more difficult home environments, this relationship is expected. What is surprising, especially in light of the biases discussed above, is how limited the impact of a woman's name is on her life outcomes once we control for other factors that are present at the time of her birth. When we include covariates in the basic specification (row 2), we find statistically significant coefficients for only four outcomes: percent Black in the hospital the mother gives birth, whether the woman is unmarried at the time of the birth, per capita income among Blacks living in her zip code, and how Black a name the woman chooses for her own child. Even in these cases of statistical significance, the magnitude of the coefficients associated with the BNI is substantively small. Changing the BNI from 50 to 100 raises the percent Black in the hospital where the mother gives birth by less than one percentage point, the probability of an unmarried birth two-tenths of one percentage point, and per capita income for Blacks in the zip code by $100. The largest impact that a woman's name appears to have is on how she names her children. An increase of 50 in the BNI raises the BNI of her children by about 3—a little more than half the impact that being a single mother has on naming, and the same impact as having the child in a hospital in which an extra 30 percent of the births are to Blacks. None of the other variables considered yield statistically significant coefficients: years of education of the woman or the father of her child, mother's age at first birth,

23. One logical solution to this problem would be to use the name of the woman's mother as an instrument. The mother's name is a strong predictor of her daughter's name, but one might plausibly argue that controlling for a wide range of covariates at the time of the daughter's birth, the mother's name will have no impact on her daughter's adult outcomes. Unfortunately, the mother's first name is not included in our data set until 1982, so this instrumental variables strategy will not be feasible until more years of time have passed.

private insurance coverage, her baby's birth weight, and number of total children born to date.[24] Particularly in light of the biases likely to be pushing the results toward finding a spurious negative relationship between names and outcomes, we conclude that there is little evidence that how Black one's name is negatively impacts life outcomes.

In the remaining rows of Table III, we expand the sample to include both Whites and Blacks. In addition to the same set of control variables included in the second row of the table, we add an indicator variable for a woman's race as well as interactions between race and BNI. The race dummy captures systematic differences in outcomes for Blacks and Whites with otherwise similar observable characteristics at birth. The interactions between BNI and race allow for a differential impact of name choice by race. Empirically, we find large coefficients on the race variable for most outcomes. Controlling for the set of characteristics observed at the time of a woman's birth, Black women live in neighborhoods with 28.8 percentage points more Blacks, give birth about one year earlier, have babies that weigh almost 200 grams less, are 26 percentage points more likely to be unmarried when they give birth, nine percentage points less likely to have private insurance, and live in neighborhoods with per capita income over $2,000 lower than Whites. The only variables for which the race dummies are not substantively large is for years of education. These systematic racial differences may reflect either discrimination or unmeasured differences between Blacks and Whites; we have no power to distinguish between these competing hypotheses. Note, however, that the weak relationship between names and outcomes persists. Interestingly, there is generally a stronger relationship between BNI and life outcomes for Whites than Blacks, although in both cases the magnitude of the effects is small.[25]

24. The findings are robust to relaxing the restriction that BNI affect the outcome measures linearly. If we replace our measure of BNI with a set of mutually exclusive indicator variables corresponding to having a unique name, a BNI greater than 80 but not unique, a BNI between 50 and 80, and the omitted category which is a BNI less than 50, the basic conclusions are unchanged.

25. We have also explored whether other aspects of naming have a causal impact on life outcomes. For example, one might expect that having the "wrong" kind of name (i.e., a White name in a Black neighborhood and vice versa) will, at least in terms of utility, adversely affect life outcomes. We attempted to identify people with the "wrong" names in two different ways. In the first approach, we calculate the deviation between the BNI that a person actually received and the BNI that we predict they should have received based on their birth circumstances. The absolute magnitude of the deviation represents how "wrong" their name is from an ex ante perspective. An alternative way of characterizing someone as

TABLE III
THE RELATIONSHIP BETWEEN NAMES AND LIFE OUTCOMES

		Percent Black in residential zip code as an adult	Percent Black in hospital where mother gives birth as an adult	Percent minority in hospital where mother gives birth as an adult	Years of education (the woman herself)	Years of education (father of woman's child)	Woman's age at first birth
	Mean	43.3	47.4	57.0	12.4	12.2	19.6
	Standard dev.	(35.0)	(27.2)	(23.3)	(1.6)	(2.2)	(3.0)
	Other controls?						
Sample restricted to Blacks							
Coefficient on the variable:							
Black Name Index	No	0.0602	0.0543	0.0458	−0.0006	−0.0006	−0.0020
		(0.0099)	(0.0080)	(0.0101)	(0.0005)	(0.0007)	(0.0008)
Black Name Index	Yes	0.0155	0.0172	0.0201	0.0008	0.0000	0.0009
		(0.0096)	(0.0072)	(0.0092)	(0.0005)	(0.0007)	(0.0007)
Sample includes Blacks and Whites							
Coefficient on the variable:							
Black Name Index * Black	Yes	0.0007	0.0076	0.0073	0.0009	0.0004	0.0003
		(0.0096)	(0.0073)	(0.0062)	(0.0005)	(0.0007)	(0.0008)
Black Name Index * White	Yes	0.0018	0.0151	0.0144	−0.0022	−0.0013	−0.0026
		(0.0028)	(0.0041)	(0.0043)	(0.00044)	(0.00066)	(0.00063)
Black	Yes	28.8	20.7	19.0	0.097	−0.072	−0.911
		(1.2)	(1.0)	(0.9)	(0.046)	(0.071)	(0.078)

TABLE III
(CONTINUED)

	Baby's birth weight in grams	Unmarried when baby born	Privately insured at time of birth	Total children born so far	Per capita zip code income (all residents)	Per capita zip code income (Black residents)	Black Name Index of name chosen for her child
	3219.4	0.71	0.37	2.01	12587	11170	77.8
	(560.4)	(0.45)	(0.48)	(1.16)	(4697)	(3552)	(29.23)
Sample restricted to Blacks							
Coefficient on the variable:							
Black Name Index	−0.27	0.0007	−0.0004	0.0009	−7.1	−5.0	0.081
	(0.16)	(0.0001)	(0.0001)	(0.0003)	(1.4)	(1.1)	(0.009)
Black Name Index	−0.04	0.0004	−0.0001	0.0005	−2.2	−1.9	0.062
	(0.17)	(0.0001)	(0.0001)	(0.0003)	(1.1)	(1.4)	(0.009)
Sample includes Blacks and Whites.							
Coefficient on the variable:							
Black Name Index * Black	0.15	0.0002	0.0000	0.0005	−0.2	−1.1	0.058
	(0.17)	(0.0001)	(0.0001)	(0.0003)	(1.4)	(1.1)	(0.009)
Black Name Index * White	0.15	0.0002	−0.0004	0.0008	−5.0	−1.8	0.034
	(0.15)	(0.0001)	(0.0001)	(0.0002)	(1.4)	(1.8)	(0.007)
Black	−194.5	0.26	−0.09	0.21	−2364.3	−687.9	41.37
	(15.1)	(0.01)	(0.02)	(0.03)	(150.7)	(132.6)	(0.82)

The dependent variable is listed at the top of each column. The values reported in the table are the coefficient on the measure of how Black an individual's name is. In the top two rows, the key regressor is the Black Name Index. In the bottom six rows, three indicator variables are included denoting whether a name is unique, greater than 80 on the Black Name Index, or between 50 and 80. The omitted category is less than 50 in those specifications. Results are presented with no controls and including percent of Black babies in birth hospital, percent of Black babies in birth county, whether mother was born in California, mother's age at time of birth, father's age at time of birth, months of prenatal care, whether born in a county hospital, birth weight, total number of children, and dummy for single mother. The sample in the top panel of the paper is the set of Black women for whom we successfully match their own birth certificate information to that of their children when the women themselves give birth. The sample in the bottom panel includes both Blacks and Whites whom we successfully match from own birth certificate to the child's. Standard errors are in parentheses. All of the covariates in the regressions (percent Black babies in woman's birth hospital, percent of Black babies in county of woman's birth, whether woman's mother is born in California, mother's and father's age at time of woman's birth, mother's months of prenatal care, whether woman's mother was married at the time of woman's birth, whether woman was born in a county hospital, woman's birth weight, and total number of prior births to the woman's mother) are known as of 1973–1974 when the women are born. All of the outcome variables correspond to the circumstances when they give birth as adults in the period 1989–2000.

There are three possible interpretations of the data that reconcile our findings with that of the audit study literature. One possibility is that the only function names serve is as a noisy initial indicator of race. In this view of the world, those with Black names may receive fewer job interviews, but no fewer job offers. Discriminatory employers will use distinctively Black names as noisy signals of race, but names become irrelevant once the candidate shows up for an interview and the employer directly observes race.[26] Indeed, in such a world, there is a benefit to Blacks of signaling their race through name choice in order to avoid undertaking costly interviews unlikely to yield jobs and to signal race to employers giving preferential treatment to Blacks. A second interpretation of the data that is consistent with our findings and those of the audit studies is that Black names, because of self-selection among Black parents, provide useful signals of human capital to employers, even controlling for race itself and other information available on resumes.[27] A final possibility is that names themselves have a modest causal impact on job callbacks and unemployment duration that we are unable to detect using the relatively coarse outcome measures available to us. Note that only in the third of these three interpretations does a person's name have any causal impact on life outcomes.

We cannot directly test between these competing hypotheses for two reasons. First, our data set lacks clean measures of productivity. We do, however, observe worker characteristics that

---

having the "wrong" name is to look at cases in which a woman's name is mismatched ex post with the actual neighborhood in which she lives as an adult. In neither case do we find a systematic link between having the "wrong" name and the types of outcomes we are measuring.

26. Even if names have an impact that extends beyond the resume stage, in an efficient market with sufficiently many nondiscriminatory employers, the presence of discrimination need not result in worse outcomes for Blacks [Becker 1957]. Note also that, at least in the past, a relatively small fraction of the jobs were actually obtained through a formal resume-based process. Granovetter [1974] reports that approximately 10 percent of respondents in a survey report obtaining their job through job advertisements.

27. From a legal perspective, the use of names as a basis for hiring decisions is likely to be a violation of current law, regardless of whether there are underlying productivity differences. In the 1971 decision *Griggs v. Duke Power,* the Supreme Court ruled unanimously that policies that are neutral on their face, but have a disparate impact on Blacks are not allowed. This doctrine was codified in the Civil Rights Act of 1991 after the 1989 case *Antonio v. Wards Cove Packing* modified the *Griggs* ruling by stating that policies with disparate impacts were allowable as long as they served a legitimate business interest. Indeed, even the use of names to choose a subset of Blacks from an entirely Black hiring pool would probably be viewed as illegal under *Connecticut v. Teal,* in which it was found that even policies that have no disparate impact on a protected group, but harm some individuals within that group, violate the law.

might be correlated with labor productivity (see, e.g., Heckman and Carneiro [2003]), but are typically not included on a resume (and are for the most part illegal for employers to inquire about): whether a woman is a single mother, was born to a teenage mother, was raised in a single-parent household, or the degree of racial segregation into which she was born. The second difficulty in testing this hypothesis is that we do not have in our data all of the information on a resume (e.g., particular schools attended, complete work history, misspellings).

Our empirical strategy, in light of these shortcomings, is to estimate a number of specifications in the general spirit of equation (2), varying the set of covariates included in the regression. A critical difference between the earlier estimates and those presented below are that we now include information about the adult circumstances of a woman as covariates, since such information is on the resume. The results are presented in Table IV. Rows of the table correspond to different factors that might be correlated with labor productivity due to their influence on human capital: segregation in one's birth hospital, mother's marital status, whether one's mother was a teenager, the woman's own birth weight, the woman's own marital status, the education level of the man who is father to her children, and the total number of children born to the woman. The first two columns of the table include both Blacks and Whites in the regressions. We do *not,* however, include race dummies in the specification because race is not directly observed by the prospective employer from the resume. The third and fourth columns of the table limit the sample only to Blacks. The entries in the table are the coefficient on the woman's BNI. Each table entry is from a different regression. The first and third columns do not include any covariates; the second and fourth column includes the woman's age, years of education, whether or not she has private health insurance (the best measure we have with respect to the quality of her current employment), zip code of residence fixed effects, and year dummies. These variables capture some of the information on a resume. Our sample is the set of women whom we observe both at their birth in 1973–1974 and as adults in 1989–2000.

The results in Table IV suggest that a woman's first name is indeed a useful predictor of the circumstances in which she grew up and her current situation, both of which may in turn be correlated with labor productivity. This is true even after including our crude set of controls that proxy for information available on the resume. A woman's name is a particularly important

TABLE IV
THE SIGNALING VALUE OF NAMES TO EMPLOYERS

Dependent variable	Sample includes Blacks and Whites		Sample includes only Blacks	
	No controls included	Controls proxying for some information on a resume	No controls included	Controls proxying for some information on a resume
Percent of Black babies in woman's birth hospital	0.3015 (0.0026)	0.1417 (0.0045)	0.0690 (0.0064)	0.0526 (0.0085)
Woman born to a teenage mother	0.00260 (0.00005)	0.00209 (0.00009)	0.00238 (0.00011)	0.00232 (0.00016)
Woman born to an unmarried mother	0.00508 (0.00005)	0.00313 (0.00009)	0.00193 (0.00011)	0.00176 (0.00017)
Woman's birth weight (in grams)	−2.16 (0.07)	−1.03 (0.09)	−0.16 (0.13)	−0.11 (0.16)
Woman herself unmarried at time she gives birth	0.00338 (0.00006)	0.00111 (0.00006)	0.00038 (0.00010)	0.00023 (0.00009)
Years of education for the man who fathers the woman's child	−0.0050 (0.0003)	0.0002 (0.0004)	−0.0010 (0.0006)	−0.0006 (0.0006)
Total number of children born to the woman	0.00283 (0.00011)	0.00238 (0.00016)	0.00079 (0.00025)	0.00090 (0.00029)

The relevant dependent variable is listed on the left-hand side of the table. The values reported in the table are the coefficients on the Black Name Index of the woman. Each table entry is from a separate regression. In columns 1 and 3, no controls are included. In columns 2 and 4, controls for a woman's age, years of education, zip code of residence, whether the woman is privately insured, and year dummies are included. These variables proxy for some of the types of information available to employers on resumes. The sample used in all regressions is the set of black women for whom we successfully link information from their own birth certificate to their child's birth certificate. Unlike earlier regressions, these specifications control for choices that a woman makes over the course of her life, such as years of education. The number of observations is 62,309 for the sample that includes Blacks and Whites and 18,427 for the sample that only includes Blacks.

predictor when we pool Blacks and Whites in the same regression (columns 1 and 2). In column 2 a woman with a BNI equal to 100 (implying a name that no Whites have) is 20.9 percentage points more likely to have been born to a teenage mother and 31.3 percentage points more likely to have been born out-of-wedlock than a Black woman living in the same zip code with the same age and education, but carrying a name that is equally common among Whites and Blacks. The woman with a Black name is also much more likely to have been born in a Black neighborhood, to herself be unmarried, to have had lower birth weight, and to have given birth to more children. The same pattern of results, although generally somewhat smaller in magnitude, is apparent even when we restrict the sample to Blacks. In other words, even if the employer knows that a candidate is Black, the Blackness of

the name continues to serve as an important signal of socioeconomic status. Thus, while we cannot rule out animus on the part of employers, we find evidence supporting a potential productivity-related statistical discrimination motive for employers to base interview decisions on first names.

V. Conclusion

We document stark differences in naming patterns between Blacks and Whites, and demonstrate that these patterns change sharply over time. While most Blacks have shifted toward more distinctively Black names (particularly in the late 1960s and early 1970s), the fraction of Blacks choosing starkly White names has also increased. How distinctively Black one's name was appears to have signaled little in the early 1960s. Naming conventions differed modestly across parental characteristics or neighborhood types. The last two decades, however, have led to a "ghettoization" of distinctively Black names, namely, a distinctively Black name is now a much stronger predictor of socioeconomic status. Among the theories we consider, models in which the rise of the Black Power movement triggers important changes in Black identity appear to be most consistent with our data. We find no relationship between how Black one's name is and life outcomes after controlling for other factors. If that conclusion is correct, then the proper interpretation of earlier audit studies using Black names on resumes is either that the impact of names does not extend beyond the callback decision (because race is directly observed at the interview stage), or that names are correlated with determinants of productivity not captured by a resume. In our data, it is difficult to distinguish between these competing hypotheses.

More generally, this paper takes first steps toward an attempt to understand what role Black culture might play in explaining continued poverty and racial isolation. With respect to this particular aspect of distinctive Black culture, we conclude that carrying a black name is primarily a *consequence* rather than a cause of poverty and segregation.

Appendix: Data Description

In this data appendix we describe the data sources that we draw upon in our analysis, give a description of some data limi-

tations and other data issues with which we are confronted, provide definitions, and describe the process used to link mothers.

1. Data Sources

Birth Data: Our empirical analyses are mainly based on data drawn from the Birth Statistical Master File maintained by the Office of Vital Records in the California Department of Health Services. These files provide information drawn from birth certificates for virtually all children born in California over the period 1961–2000. With the approval of the California Committee for the Protection of Human Subjects, we have been able to supplement the information contained on the public use versions of these data sets with personal identifiers including mother's first name, mother's maiden name, and child's full name.

Zip Code Statistics: U. S. Bureau of the Census, American Fact Finder, 1990 Census.

Demographic and Socioeconomic Data: For an accuracy check on mother-children links, data were obtained from the American Community Survey 5 Percent Public Use of Microdata Samples (PUMS) for California and United States (All States), U. S. Bureau of the Census *Arabic First Names:* To identify and discard Arabic names from our data, we used Bruce Lansky [1999], *Baby Names Around the World.*

2. Data Limitations and Other Data Issues

Race/Ethnicity: From 1998 to 2000, information of a child's race is missing. For these three years, a child is considered Black if either parent is Black. In other years, over 98 percent of children with at least one Black parent are coded as Black. Prior to 1982, most Hispanics are included with whites, and it is not possible to directly calculate statistics for Non-Hispanics. Beginning in 1982, an identifier for Hispanic ethnicity was added to the birth certificate in California. Using information from this latter period, we drop anyone in the earlier time period with a surname or mother's maiden name that is more than 10 percent Hispanic.

Link of Birth Certificate Data for Women Born in 1973–1974 to When They Give Birth Themselves in the Period 1989–2000: In order to uniquely link information from a woman's own birth certificate to those of her children when she gives birth, a woman must both be born in California and later give birth there. Only after 1989 do we have the mother's date of birth and full first name, two critical pieces of information for making the linkage. We thus limit the sample to babies born 1989–2000 to Black women themselves born in California. In order to identify women

whose birth histories are likely to be most fully captured by the period 1989–2000, we restrict the matching to women who were themselves born in 1973–1974. We restrict the sample to Black women born in California. We then carry out five different matching procedures in which we use a mother's exact date of birth and permutations of her first name and maiden name (match only on maiden name, only on first name, on first name and maiden name, on first initial and maiden name, and on first name and maiden name initial). We keep only cases where the match between mother and child is unique; i.e., there is no other woman born in California in 1973–1974 who has information that could possibly match the characteristics of the mother on the birth certificate. Because of typographical errors, using only a subset of the information in a woman's name somewhat increases the number of successful matches.

HARVARD UNIVERSITY SOCIETY OF FELLOWS AND NBER
DEPARTMENT OF ECONOMICS, UNIVERSITY OF CHICAGO

REFERENCES

Akerlof, George, and Rachel Kranton, "Economics and Identity," *Quarterly Journal of Economics,* CXV (2000), 715–753.
Anderson, Elijah, *Streetwise: Race, Class, and Change in an Urban Community* (Chicago, IL: University of Chicago Press, 1990).
——, *Code of the Street: Decency, Violence, and the Moral Life of the Inner City* (New York, NY: W. W. Norton and Company, Inc., 1999).
Austen-Smith, David, and Roland Fryer, "The Economics of 'Acting White'," NBER Working Paper No. 9904, 2003.
Bart, Barbara, Marsha Hass, Jane Philbrick, Martha Sparks, and Craig Williams, "What's in a Name," *Women in Management Review,* XII (1997), 209–308.
Becker, Gary, *The Economics of Discrimination* (Chicago, IL: University of Chicago Press, 1957).
Berman, Eli, "Sect, Subsidy, and Sacrifice: An Economist's View of Ultra-Orthodox Jews," *Quarterly Journal of Economics,* CXV (2000), 905–953.
Bertrand, Marianne, and Sendhil Mullainathan, "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment Evidence on Labor Market Discrimination," NBER Working Paper No. 9873, 2003.
Bound, John, and Richard Freeman, "What Went Wrong? The Erosion of Relative Earnings and Employment among Black Men in the 1980s," *Quarterly Journal of Economics,* CVII (1992), 201–232.
Brown, C., and P. Gay, *Racial Discrimination 17 Years After the Act* (London, UK: Policy Studies Institute, 1985).
Busse, Thomas V., and Louisa Seraydarian, "Desirability of First Names, Ethnicity and Parental Education," *Psychological Reports,* XL (1977), 739–742.
Chandra, Amitabh, "Is the Convergence of the Racial Wage Gap Illusory?" NBER Working Paper No. 9476, 2003.
Counihan, Carol, and Penny Van Esterik, *Food and Culture, A Reader* (New York, NY: Routledge, 1997).
Cutler, David, Edward Glaeser, and Jake Vigdor, "The Rise and Decline of the American Ghetto," *Journal of Political Economy,* CVII (1999), 455–506.
Fordham, Signithia, and John Ogbu, "Black Students' School Successes: Coping with the Burden of 'Acting White'," *The Urban Review,* XVIII (1986), 176–206.

Fryer, Roland G., "An Economic Approach to Cultural Capital," unpublished manuscript, University of Chicago, 2003.
Fryer, Roland, and Steven Levitt, "The Causes and Consequences of Distinctively Black Names," unpublished manuscript, University of Chicago, 2002.
Glaeser, Edward, David Laibson, and Bruce Sacerdote, "An Economic Approach to Social Capital," *Economic Journal,* CXII (2002), 437–458.
Goldin, Claudia, and Maria Shim, "Making a Name: Women's Surnames at Marriage and Beyond," mimeo, Harvard University, 2003.
Granovetter, Mark, *Getting a Job: A Study of Contacts and Careers* (Chicago, IL: University of Chicago Press, 1974).
Hannerz, Ulf, *Soulside: Inquiries into Ghetto Culture and Community* (New York, NY: Columbia University Press, 1969).
Heckman, James, Tom Lyons, and Petra Todd, "Understanding Black-White Wage Differentials, 1960–1990," *American Economic Review,* XC (2000), 344–349.
Heckman, James, and Pedro Carneiro, "Human Capital Policy," NBER Working Paper No. 9495, 2003.
Hoffman, Charles, "Historically Black Colleges and Universities, 1976–1994," NCES Technical Paper No. 96902, 1996.
Hubbick, J., and S. Carter, *Half a Chance? A Report on Job Discrimination against Young Black Males in Nottingham* (London, UK: Commission for Racial Equality, 1980).
Iannaccone, Laurence, "Sacrifice and Stigma: Reducing Free-Riding in Cults, Communes, and Other Collectives," *Journal of Political Economy,* C (1992), 271–291.
Jackson, George, *White and Negro Spirituals, their Life Span and Kinship, Tracing 200 Years of Untrammeled Song Making and Singing among our Country Folk, with 116 Songs as Sung by Both Races* (New York, NY: J. J. Augustin, 1944).
Jencks, Christopher, and Meredith Phillips, *The Black-White Test Score Gap* (Washington, DC: Brookings Institution, 1998).
Johnston, Lloyd, Patrick O'Malley, Jerald Bachman, and John Schulenberg, "Cigarette Brand Preferences among Adolescents," Monitoring the Future Occasional Paper 45, Institute for Social Research, University of Michigan, 1999.
Jowell, R., and P. Prescott-Clarke, "Racial Discrimination and White-Collar Workers in Britain," *Race,* XI (1970), 397–417.
Kington, Raynard, and Herbert Nickens, "Racial and Ethnic Differences in Health: Recent Trends, Current Patterns, and Future Directions," in *America Becoming: Racial Trends and Their Consequences,* Neil J. Smelser, William J. Wilson, and Faith Mitchell, editors, Commission on Behavioral and Social Sciences and Education (Washington, DC: National Academy Press, 2001), pp. 253–310.
Lazear, Edward, "Culture and Language," *Journal of Political Economy,* CVII (1999), S95–S126.
Lewis, Oscar, *La Vida: A Puerto Rican Family in the Culture of Poverty—San Juan and New York* (New York, NY: Random House, 1966).
Lieberson, Stanley, *A Matter of Taste: How Names, Fashions, and Culture Change* (New Haven, CT: Yale University Press, 2000).
Lieberson, Stanley, and Kelly Mikelson, "Distinctive African-American Names: An Experimental, Historical, and Linguistic Analysis of Innovation," *American Sociological Review,* LX (1995), 928–946.
Massey, Douglas, "Residential Segregation and Neighborhood Conditions in U. S. Metropolitan Areas," in *America Becoming: Racial Trends and Their Consequences,* I (Washington, DC: The National Academies Press, 2001), pp. 391–434.
McAdams, Doug, Roland Fryer, and Steven Levitt, "What's in a Name? Black Power and the Change in African-American Names, 1961–2000," working paper, Stanford University, 2003.
Orr, Eleanor, *Twice as Less: Black English and the Performance of Black Students in Mathematics and Science* (New York, NY: W. W. Norton, 1989).
Reno, Janet, et al., *Correctional Populations of the United States* (Washington, DC: Bureau of Justice Statistics, 1997).
Riessman, Frank, *The Culturally Deprived Child* (New York, NY: Harper & Bros, 1962).

Smith, James, and Finis Welch, "Black Economic Progress After Myrdal," *Journal of Economic Literature,* XXVII (1989), 519–564.
Van Deburg, W. L., *New Day in Babylon: The Black Power Movement and American Culture, 1965–1975* (Chicago, IL: University of Chicago Press, 1992).
Ventura, Stephanie, and Christine Bachrach, "Nonmarital Childbearing in the United States, 1940–1999," in *National Vital Statistics Reports,* XLVIII (Hyattsville, MD: National Center for Health Statistics, 2000).
Waldfogel, Joel, "Preference Externalities: An Empirical Study of Who Benefits Whom in Differentiated Product Markets," *Rand Journal of Economics,* XXXIV (2003), 557–568.
Wilson, William, *The Truly Disadvantaged: The Inner City, the Underclass and Public Policy* (Chicago, IL: University of Chicago Press, 1987).
Wolfram, Walt, and Erik Thomas, *Development of African-American English* (Oxford, UK: Blackwell Publishers, 2002).
Woodward, Vann C., *The Strange Career of Jim Crow* (New York, NY: Oxford University Press, 1974).
Young, Andrew, *An Easy Burden: The Civil Rights Movement and the Transformation of America* (New York, NY: HarperCollins Publishers, 1996).

E-lit-004: Systemic Discrimination: Theory and Measurement

J. Aislinn Bohren, Peter Hull and Alex Imas | Quarterly Journal of Economics 140(3): 1743-1799 (NBER Working Paper 29820, March 2022, revised December 2024) | 2025-08-01

https://www.nber.org/system/files/working_papers/w29820/w29820.pdf

About this source

The paper's contribution to this section is the direction of the bias in a Bertrand-Mullainathan-style estimate: holding the resume fixed identifies the causal effect of the race signal at that one hiring decision, and therefore excludes any disparity that arose because discrimination elsewhere in the system altered what is on the resume. Their two-node field experiment makes that concrete. In a national sample of automotive firms - the sector Kline, Rose and Walters had flagged as the most discriminatory - resumes with no prior work experience showed a 12.4 percentage point Black-White callback gap (25.6 versus 13.2 percent), larger than the canonical estimates, which the authors attribute to those studies having put experience on every resume. Adding one line of relevant experience shrank the gap to a statistically insignificant 4.2 points, close to Kline et al.'s 5.3-point automotive gap, while experience itself raised callbacks by 9.5 points (49 percent). Because early discrimination denies Black applicants the experience that later employers reward, direct discrimination at the first node feeds forward: at the median local market thickness of 8 openings the second-node White-Black hiring gap is about 7.5 percentage points, and more than half of it is systemic rather than direct. The authors conclude that a standard correspondence study, which conditions on experience, would miss nearly half of the total discrimination in this two-node system - and more if experience is shaped at more than two nodes. Their gender experiment pushes the point further: with recommendation letters held fixed, gender gaps in hiring likelihood and wage are insignificant, but letting the letters be generated from the applicant's own gendered resume produces a 21 percent wage gap, so essentially all measured gender discrimination is systemic and invisible to a design that holds materials fixed. The paper offers no evidence bearing on the separate criticism that distinctively Black names also signal socioeconomic status; it takes the name manipulation from the literature and uses it unchanged.

Bound as the original PDF. Retrieved 2026-09-10. sha256 0973ae89c4763ef0...

Bound into an Economic Briefing · by Claude, may contain mistakes

Full document begins on the next page.

NBER WORKING PAPER SERIES

SYSTEMIC DISCRIMINATION: THEORY AND MEASUREMENT

J. Aislinn Bohren
Peter Hull
Alex Imas

Working Paper 29820
http://www.nber.org/papers/w29820

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2022, Revised December 2024

We thank Jason Abaluck, Amanda Agan, David Arnold, Marianne Bertrand, Peter Blair, Leonardo Bursztyn, Hanming Fang, Damon Jones, Diag Davenport, Stefano DellaVigna, Will Dobbie, Ed Glaeser, Sam Hirshman, Larry Katz, Erzo Luttmer, Matthew Knepper, Matt Lowe, Muriel Niederle, Matthew Notowidigdo, Anna Gifty Opoku-Agyeman, Jane Risen, Evan Rose, Heather Sarsons, Andrei Shleifer, Fabio Tufano, Dami an Vergara, Basit Zafar, and seminar participants and conference participants at various institutions for helpful comments. Cuimin Ba, Joshua Hascher, Christy Kang, Matthew Murphy and Yier Ling provided expert research assistance. Hull gratefully acknowledges funding through the National Science Foundation (Award #2119849). 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.

© 2022 by J. Aislinn Bohren, Peter Hull, and Alex Imas. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Systemic Discrimination: Theory and Measurement
J. Aislinn Bohren, Peter Hull, and Alex Imas
NBER Working Paper No. 29820
March 2022, Revised December 2024
JEL No. D63, D83, J16, J71

ABSTRACT

Economists often measure discrimination as disparities arising from the direct effects of group identity. We develop new tools to model and measure systemic discrimination, which instead captures how discrimination in other decisions indirectly contributes to disparities. We propose an experimental design, the Iterated Audit, to identify systemic discrimination. We then illustrate these new tools in two field experiments. The first experiment shows how racial discrimination accumulates across multiple rounds of hiring through the interaction of two forces: greater discrimination against inexperienced workers—which affects the opportunity to obtain experience—and high subsequent returns to experience. The second experiment shows how gender-based differences in the language of recommendation letters can translate into systemic gender discrimination in STEM hiring. We discuss how our findings qualify previous results on direct discrimination and outline how our tools can be used to target policy interventions.

J. Aislinn Bohren
Department of Economics
The Ronald O. Perelman Center for
Political Science and Economics
University of Pennsylvania
133 South 36th Street
Philadelphia, PA 19104
abohren@gmail.com

Alex Imas
Booth School of Business
University of Chicago
5897 S. Woodlawn Avenue
Chicago, IL 60637
and NBER
alex.imas@chicagobooth.edu

Peter Hull
Department of Economics
Box B, Brown University
Providence RI 02912
and NBER
peter_hull@brown.edu

1 Introduction

Disparities by race, gender, and other protected characteristics are widely documented, raising concerns of discrimination. In economics, such concerns are usually probed with a rich set of tools for modeling and measuring *direct* discrimination: how protected characteristics affect individual actions, holding fixed all other relevant factors. An economist might, for example, measure direct discrimination by estimating the causal effect of a hiring manager's perceptions of a job applicant's race, holding fixed the applicant's work experience and education. She might interpret any such race effects through canonical models of taste-based or statistical discrimination, as well as other theories of direct discrimination.

There is, however, a growing recognition that focusing on direct discrimination can yield an incomplete understanding of how societal inequities can arise, persist, and compound. Sociologists and legal scholars have long emphasized the importance of systems-based analyses, which study discrimination as the cumulative outcome of interactions across different periods and domains (Pincus 1996; De Plevitz 2007; Powell 2008; Small and Pager 2020). Decades of research in labor economics similarly notes how "pre-market" discrimination in education and housing might affect the employment opportunities of minorities even in the absence of direct discrimination (Cain 1986; Neal and Johnson 1996; Bertrand and Mullainathan 2004; Blank 2005).[1] More recently, computer scientists have shown how discrimination in algorithmic decisions can arise indirectly from biased data collection and training even when the algorithm is "blinded" to protected characteristics (Angwin et al. 2016; Rambachan and Roth 2020). But despite these insights, when compared with the robust toolkit for modeling and measuring direct discrimination, economists have more limited theoretical and empirical tools to study such indirect or *systemic* forms of discrimination (Small and Pager 2020).

This paper develops a common theoretical framework for studying direct and systemic discrimination, and new empirical tools for bringing this framework to data. We focus on a notion of systemic discrimination that captures how direct discrimination in other decisions leads to differences in relevant attributes for a given decision, which in turn generates disparities in outcomes. Our framework contributes to the sociology and legal literatures a precise mathematical language for analyzing such systemic discrimination and its drivers.[2] To the

[1] For example, Cain (1986) presents two statistical models for measuring discrimination: model (I) identifies discrimination as group-based differences in an outcome variable of interest, controlling for all relevant productivity characteristics; model (II) identifies discrimination as the unconditional difference in the outcome variable and implicitly attributes any differences in productivity characteristics to pre-market discrimination. Our framework provides a decomposition of this unconditional difference into different forms of systemic discrimination as well as potentially non-discriminatory disparities, and allows for measurement of the different systemic channels.

[2] As Small and Pager (2020) note, terms like "systemic" or "structural" discrimination—while broadly referring to the idea that "something other than individuals may discriminate"—are often imprecisely and inconsistently used across the social sciences.

labor economics literature on pre-market discrimination and the computer science literature on algorithmic fairness, we contribute a general approach for formally modeling systemic discrimination in a wide range of settings. Adding structure to the notion of pre-market discrimination yields a novel empirical strategy—the *Iterated Audit*—that builds on existing experimental methods to measure both forms of discrimination within a pre-defined system. We show how this approach can be used to quantify the impact of different systemic factors on observed disparities and to help inform potential policy responses.

We start by developing a general framework for examining direct and systemic discrimination, including how the former can contribute to the latter through interactions across time and domains. We model a "system" as a network of interconnected nodes, each representing a decision (e.g., a hiring manager considering the individual for a job) that can affect relevant attributes (e.g., work experience) at other decision nodes. Within this system, disparities at a given 'focal' node arise from three distinct channels: direct discrimination at that node (e.g., preferring to hire men over women with identical experience), systemic discrimination arising from the impact of direct discrimination at other nodes (e.g., direct discrimination in past hiring leading to differential experience for women vs. men, and hence further hiring disparities), and differences in characteristics that do not arise from discrimination (e.g., gendered differences in innate physical capacity for work). Direct discrimination at other nodes contributes to systemic discrimination at the focal node when it generates differences in the signaling technology (e.g., experience signals productivity) or payoff-relevant characteristics (e.g., experience increases productivity itself), which lead to treatment disparities at the focal node (e.g., experienced workers are more likely to be hired).

This framework highlights a key analytic choice that researchers interested in quantifying systemic discrimination must make: which systemic forces to analyze. Researchers may wish to hone in on a subset of nodes in order to isolate how decisions within this subset contribute to disparities at the focal node—perhaps due to the availability of detailed decision data for these nodes, policy-makers' capacity to intervene at these nodes, or simply because these are the systemic forces of interest. This amounts to a choice of decision nodes to include in the analysis, which we refer to as the *subsystem*, and a measure to group comparable workers at entry into this subsystem, which we refer to as the *reference qualification*.

For example, a researcher interested in studying systemic discrimination in hiring (the focal node) stemming from direct discrimination by reference letter writers would include both the hiring and reference letter nodes in the subsystem. This would allow her to measure how direct discrimination in reference letters translates to disparities in entry-level hiring. Another researcher interested in the impact of discrimination in both internship opportunities and letter writing on hiring would study a 3-node subsystem that also includes the internship node. Finally, a third researcher interested in studying all informational sources of systemic

discrimination would include all nodes that generate disparities in the information available at the hiring node for workers with the same hiring-node productivity. These approaches contrast with a researcher only interested in direct discrimination in hiring, who would include only this node in the subsystem. What we refer to as *total* discrimination aggregates disparities at the focal node stemming from systemic discrimination within the subsystem and any focal-node direct discrimination.[3] Thus, by choosing different subsystems, the framework breaks down the complex system that generates disparities into interpretable and measurable components and gives a unified structure for different notions of discrimination.[4]

We next show how this framework can be brought to data with the Iterated Audit (IA) approach. An IA analysis involves iterating decisions across multiple nodes. It has two components: (i) a treatment component capturing direct discrimination at the focal node and other included nodes, and (ii) an interaction component capturing how actions at the other included nodes impact treatment at the focal node. Consider, for example, a two-node subsystem in which individuals first apply for an internship and then an entry-level position (the focal node). An IA analysis involves measuring (i) direct discrimination in internship and entry-level hiring, and (ii) how internship experience impacts entry-level hiring for individuals with otherwise similar resumes. The latter interaction component shapes how direct discrimination in internship hiring drives systemic discrimination in entry-level hiring.

Direct discrimination, and hence the IA treatment component, can be identified by conventional experimental designs (e.g., audit or correspondence studies). We develop two identification strategies for the interaction component, and hence systemic discrimination. The *constructive* IA approach separately estimates the impact of each possible action at another node on treatment at the focal node; combining these estimates identifies the interaction component. The *experimental* IA approach directly measures systemic discrimination by simulating the distribution of focal-node signals, as impacted by group membership and the actions at other nodes, and measuring focal-node actions given the simulated signals. For example, the interaction component above could be constructed from estimates of how having internship experience affects entry-level hiring rates by race for individuals with otherwise similar resumes. Alternatively, one could simulate internship experience by race and measure subsequent entry-level hiring. The latter experimental approach may be preferred when attributes are high-dimensional or otherwise complex (e.g., text data), making application of

[3]Any remaining disparity at the focal node may either stem from systemic discrimination arising from nodes outside of the subsystem or non-discriminatory factors.

[4]Most analyses of systemic discrimination in sociology can be understood as choosing a measure that includes all possible nodes. This also corresponds to model (II) in Cain (1986). As the author notes, however, such measures captures both obvious discrimination and what might be viewed as non-discriminatory differences in attributes (e.g., innate physical characteristics). This paper shows how more structured measures of systemic discrimination can distinguish between these two components, as well as separates the first component into direct and systemic components.

the constructive approach difficult. Both approaches improve over simpler alternatives—such as those that simply add conventional direct discrimination estimates across nodes without measuring the interaction component—particularly when actions at other nodes impact decisions at the focal node non-linearly or when decision-makers act to undo any perceived direct discrimination at other nodes.

We illustrate these new tools in two field experiments. The first experiment used the constructive IA approach to study how direct racial discrimination in entry-level job hiring can generate systemic discrimination in later hiring via disparities in applicant work experience.[5] Here, we considered a system with two nodes. Applicants apply for jobs without prior work experience in the first round of hiring (the first node) and either obtain a job or not. They then apply for a job in the second round of hiring (the second, focal node), with or without prior experience from the first round. Direct discrimination in first-round hiring can thus become embedded in work experience, leading to systemic discrimination in second-round hiring. We used a correspondence study to estimate direct discrimination in both rounds, then used additional data to construct the interaction component and estimate systemic discrimination.

Specifically, we built on the correspondence study methodology (Bertrand and Mullainathan 2004; Kline et al. 2022) by generating job applications for a fictitious group of workers that vary in their level of experience and submitted them to online job vacancies at a set of national firms. We focused on the automotive firms which Kline et al. (2022) document as having the highest levels of direct discrimination in callback rates, and similarly randomize applicants' names to signal their race.[6] In the first round of hiring, we found sizable direct discrimination: among those with no previous work experience, applicants with distinctively White names were 13 percentage points (90%) more likely to receive a callback than applicants with distinctively Black names.[7]

Estimating the interaction component, and hence systemic discrimination, requires assessing (i) the return to experience in the second round of hiring and (ii) how direct discrimination in the first round translates to race-based differences in experience. For the former, we sent out resumes that had one line of previous experience—at similar firms as the target job—which differed only in the name of the applicant. We found substantial returns

[5]The impact of entry-level disparities on later job-market outcomes is highlighted in the 2017 statement by the Association of Women Surgeons: "The disparities women face in compensation at entry level positions lead to a persistent trend of unequal pay for equal work throughout the course of their careers."

[6]Kline et al. (2022) ran a correspondence study across a large set of US firms. The study used fictitious resumes that were identical except for the applicant's name, which was either distinctly Black or distinctly White. Importantly, each resume had at least some level of experience. The authors observed a significant 5.3 percentage-point callback gap between Black and White resumes sent to automotive firms.

[7]This gap is higher than previous estimates of callback effects which, to the best of our knowledge, used resumes with prior experience (Bertrand and Mullainathan 2004; Nunley et al. 2015; Deming et al. 2016).

to experience: overall, applicants with one line of previous experience were 10 percentage points (50%) more likely to receive a callback than those without. This suggests a meaningful role for systemic discrimination, as direct discrimination in the first round affects an important attribute (experience) for second-round hiring. How this direct discrimination translates depends on local market thickness—i.e., the number of jobs a first-round applicant can apply to—and the rate at which first-round callbacks convert to employment. To estimate market thickness, we scraped the number of job openings across the municipalities in our experiment. Finally, we estimated callback conversion rates by surveying a separate sample of hiring managers in the automotive industry.

Combining these data, we found significant systemic discrimination in second-round hiring—comprising roughly half of the measured total discrimination. The other half was due to direct discrimination, which was lower in the second round due to lower direct discrimination against experienced applicants (4 percentage points compared to the 13 percentage points documented for inexperienced applicants). This implies that simply looking at the conventional direct measure would have underestimated total discrimination by about 50%, highlighting the importance of these new tools for measuring the full extent of racial inequity. Our results also illustrate the utility of the IA method for assessing potential policy responses. The size of second-round systemic discrimination suggests that targeting first-round direct discrimination would significantly mitigate disparities in subsequent rounds. Moreover, the role of local market thickness in shaping systemic discrimination suggests scope for further policy targeting: systemic discrimination is lower in markets with low or high levels of thickness, implying that policy directed to areas with intermediate levels of thickness would be most effective for mitigating the compounding impacts of direct discrimination.

Our second lab-in-the-field experiment demonstrates the value of the experimental IA method in settings with high-dimensional or complex signals. In this study, we measured how direct gender discrimination in the language of recommendation letters can generate systemic discrimination in hiring. As before, we considered a system with two nodes. Applicants receive recommendation letters based on their resumes (the first node), then apply for a job (the second, focal node) by submitting their resumes and recommendation letters. Prior work has found significant language differences in the letters of similarly qualified male and female applicants (e.g., Schmader et al. 2007), which are replicated in automated recommendation letters from large language models (LLMs) (Wan et al. 2023). Direct discrimination in recommendation letter language thus becomes a part of the job-seekers' application materials. But it is not obvious how such language differences may contribute to subsequent hiring disparities: hiring managers may focus on "hard" information contained in the resume and ignore the differences in recommendation letters, or these differences may not lie in language that is most relevant for the hiring decisions. Estimating these interactions is challenging

because of the high-dimensionality of text data.

To quantify the systemic impact of recommendation language disparities on labor market outcomes, we randomized distinctively male or female names across a set of fictitious resumes and generated recommendation letters from them via standard LLM-based techniques. As in prior work, the resulting recommendation letters displayed marked gender-based differences in language. Following the experimental version of the IA method, we then generated three sets of "materials" (resumes and recommendation letters). The first two sets (A and B) followed a standard correspondence study design, with set A assigned distinctively male names, set B assigned distinctively female names, and the recommendation letters in both sets corresponded to those originally generated for male candidates. The third set, C, was assigned female names and recommendation letters generated for female candidates. We submitted these materials to a set of real-world hiring managers and elicited expected hiring probabilities and wages via an incentivized ratings design (Kessler et al. 2019). A comparison of outcomes for set A vs. B thus identifies direct gender discrimination, holding fixed applicant materials, while a comparison of B vs. C identifies systemic discrimination from the direct discrimination in recommendation letters. Taken together, a comparison of A vs. C identifies total discrimination inclusive of systemic language disparities in the letters.

We found that essentially all of the gender discrimination in hiring rates and wages was driven by systemic disparities from recommendation letter language. Overall, applicants with distinctively male names and male recommendation letters were substantially more likely to be hired and were assigned a 21% higher wage than applicants with distinctively female names and female recommendation letters. Holding recommendation letters fixed, however, shrinks the hiring and wage disparities to insignificant levels.

These findings are in line with recent work on gender discrimination in labor markets which suggests a limited role of direct discrimination in explaining observed gender disparities. For example, a recent meta-analysis of correspondence studies in Schaerer et al. (2023) found little discrimination in aggregate when non-group attributes are held fixed (though there was substantial heterogeneity across employers/sectors, with some favoring women and others favoring men; see also Kline et al. 2022 and Ceci and Williams 2011). Our results suggest that systemic discrimination arising from direct discrimination in how male and female candidates are described to employers can potentially explain some of the total disparities observed in the labor market that conventional direct discrimination measures miss.

This paper builds on several related literatures in economics. The notion of pre-market discrimination in labor economics is a form of systemic discrimination in our framework, and some models in this literature (e.g., Coate and Loury 1993; Cornell and Welch 1996) microfound a particular system of nodes. Our theoretical framework nests these models, capturing both broader notions of systemic discrimination from outside of economics (e.g.,

Gynter 2003; Feagin 2013) as well as more modern notions of indirect economic discrimination (e.g., Hurst et al. 2024). Connecting these different literatures yields a general approach to measurement and a unified framework for considering appropriate policy responses. More recently, McMillon (2024) provides a taxonomy to delineate different ways an initial instance of discrimination can propagate within a discriminatory system. He also suggests that such systems may be exploited to amplify the effects of equity-focused interventions.

Empirically, our approach relates to a concern from the economics discrimination literature of "bad controls"—i.e., conditioning on characteristics that are themselves affected by discrimination (e.g., Cain 1986; Altonji and Blank 1999). Indeed, the question of what to control for in a regression of outcomes on group membership is similar to the choice of subsystem and reference qualification. This choice leads naturally to the specification of "good" and "bad" controls in our framework. Given a set of "good" controls, our framework also decomposes the resulting conditional disparity into direct and systemic components, and directly links the systemic component to direct discrimination at particular other decision nodes. Discrimination measures with more and fewer controls can thus be reconciled as identifying different forms of systemic and total discrimination.

Finally, our experimental findings add to the growing literature estimating the impact of previous direct discrimination on subsequent disparities (Cook 2014; Williams et al. 2021; Derenoncourt et al. 2024; Eli et al. 2023; Harrington and Shaffer 2023). A series of recent empirical papers also build directly on our framework to measure and classify direct, systemic, and total discrimination in various settings (Zivin and Singer 2022; Lodermeier 2023; Gawai and Foltz 2023; Buchmann et al. 2023; Abramitzky et al. 2023; Althoff and Reichardt 2024; Baron et al. 2024). More broadly, we join a literature modeling and estimating the indirect impact of discrimination on important economic outcomes—including Darity (2005), Bohren et al. (2019), Arnold et al. (2022), Bohren et al. (2023), and Hurst et al. (2024).

We organize the remainder of this paper as follows. Section 2 presents a simple example motivating a systems-based approach to discrimination. Section 3 develops our theoretical framework for studying direct and systemic discrimination. Section 4 discusses measurement, including the IA design. Sections 5 and 6 present the empirical applications. Section 7 concludes. Appendix A reviews connections to related literatures, Appendix B presents an additional application, and Appendix C presents further details of the empirical studies.

2 Motivating Example

We start with a simple example that illustrates key features of the framework. Consider a population of computer programmers applying to a job at a software company to code a new feature. Let $Y_i^* \in \{0, 1\}$ indicate whether, if hired, programmer i would be able to successfully code the feature and let $A_i^* \in \{0, 1\}$ indicate whether programmer i is hired.

To hire programmers, the company solicits a resume R_i and a performance profile P_i for each applicant, where P_i is a rating of performance on coding assignments the applicant completed on an online coding platform such as GitHub. Let C_i denote the code i submitted to the platform, which forms the basis of his or her rating P_i.[8] The company's hiring manager observes this resume and performance profile for each programmer i, denoted by $S_i^* = \{R_i, P_i\}$, as well as the programmer's gender G_i.

A large literature documents gender-based disparities in hiring in STEM fields (e.g., Davison and Burke 2000; Roth et al. 2012). Suppose three economists are interested in studying the role of discrimination in generating such disparities. Economist 1 follows standard practices in the field with a carefully designed correspondence study. Specifically, she sends the company a set of fictitious resumes and performance profiles $S_i^* = \{R_i, P_i\}$, along with a signal of each fictitious applicant's gender G_i (e.g., a distinctively male or female first name), and elicits a hiring decision A_i^*. By randomizing G_i, she is able to estimate the causal effect of gender on hiring decisions conditional on the resume and profile S_i^*. She finds that male applicants are slightly less likely to be hired than female applicants with the same resume and profile. She concludes that there is some discrimination against male applicants.

Economist 2 is interested in the same question, but ends up running a slightly different correspondence study. Rather than generating fictitious performance profiles, she generates code for each fictitious applicant and submits it to the *online platform* (along with the salient signal of the applicant's gender). The performance profile is generated by evaluators on the platform based on this code. The economist then submits fictitious resumes and these performance profiles to the company. Thus, while applicants have comparable resumes and code, their performance profiles may differ based on how they are evaluated on the platform. Strikingly, she reaches a different conclusion than Economist 1: male applicants are slightly *more* likely to be hired than female applicants with the same resume and submitted code. She concludes there is some discrimination against female applicants in this setting. In unpacking this result, she finds that evaluators on the platform tend to be less willing to accept the same code from female programmers than male programmers, which generates gender-based disparities in the resulting profiles.[9]

Finally, Economist 3 examines the same question by running a different type of study. She recruits a real set of male and female programmers with similar resumes and performance

[8] Programmers' contributions to Open Source Software (OSS) projects often play a role in labor market hiring decisions. OSS projects are typically hosted by project managers on platforms such as GitHub, where project managers post 'pull requests' which allow other users to contribute code for specific aspects of the project. The project manager evaluates this code and decides to accept or reject it. A user's history of accepted pull requests is part of their GitHub profile, which is often considered in hiring decisions. Prior work has documented gender-based disparities in the acceptance of pull requests on GitHub (Terrell et al. 2017), as well as in performance profiles on other coding platforms, e.g., Stack Overflow (May et al. 2019).

[9] See Terrell et al. (2017) for evidence of such disparities.

profiles as the fictitious applicants in the first two studies. She then hires them to code a new feature similar to the one desired by the company. In this way, Economist 3 is able to measure the coding ability Y_i^* of each programmer. Submitting these programmers' resumes and performance profiles to the company, she computes gender disparities in elicited hiring decisions among applicants with the same true coding ability—without conditioning on the resume, code submitted to the platform, or performance profile. Curiously, she reaches a different conclusion than both Economist 1 and 2: male computer programmers are *much* more likely to be hired and receive a positive evaluation than female computer programmers with the same underlying ability. In unpacking this result, Economist 3 finds that a key driver is differential prior experience: among programmers with the same coding ability, female programmers were hired for far fewer coding jobs.

At first blush, this simple example presents a puzzle: which of the three researchers are correct on the nature and extent of discrimination in STEM hiring? Economist 1 follows the norm in audit and correspondence studies by conditioning on the information S_i^* available to the decision-maker of interest (i.e., the software company). Economist 2 finds that some of this signal is biased by a different evaluation (i.e., the performance profile); her measure of discrimination takes this bias into account. By conditioning on an "objective" measure of coding ability Y_i^*, Economist 3 uncovers a further bias in the applicant's opportunity to gain experience. How can the *systemic* biases that Economists 2 and 3 find be studied alongside the *direct* discrimination Economist 1 finds?

In Section 3 we develop a general framework that reconciles these analyses. The framework formalizes how the study of discrimination beyond a single decision requires a researcher to select a set of other decisions to include in the analysis—what we call the subsystem—which is the key factor that differed across the three economists here. In any given setting, by selecting different subsystems, a researcher can study different systemic forces alongside canonical sources of direct discrimination.

Two other points, which we return to in subsequent sections, are worth highlighting in this example. First, as shown in Section 4, studying such systemic forms generally requires new empirical tools. While Economist 1 identified direct discrimination with a standard correspondence study, isolating the effects of the systemic biases found by Economists 2 and 3 is more challenging. We propose an new Iterated Audit design to identify these effects.

Second, this example highlights how it may be difficult to address such systemic forms of discrimination with standard individual-level interventions. The hiring managers at the software company in this example seem at least partly aware of the systemic bias in performance profiles, given their "reverse" discrimination in evaluating female versus male applicants. Yet they do not fully offset this bias, either because of imperfect awareness, psychological frictions, or their own biases. Hence, broader or system-wide policy responses may be called

for in settings with significant systemic discrimination. By nesting different forms of discrimination in a single framework, our approach can be used to formulate and target such system-wide policy responses and study how they may impact other interconnected decisions.

3 A Framework for Systemic Discrimination

We now develop a theoretical framework to formalize a notion of systemic discrimination. Section 3.1 introduces the setting and defines three forms of discrimination: direct, total, and systemic. Section 3.2 relates these via a decomposition of total discrimination into direct and systemic components. Section 3.3 discusses key features of the framework: how it relates to notions of systemic discrimination in other fields, how it nests existing notions of discrimination in the economics literature through the choice of a subsystem, the sources of systemic discrimination, and the relationship between systemic and statistical discrimination.

3.1 Framework and Definitions

We develop the framework in a labor market context, in which we consider discrimination among workers being evaluated for a task n^*. Worker i has ex-ante unobservable *productivity* $Y_i^* \in \mathcal{Y}^*$ on the task (e.g., error rate, number of units produced), where $\mathcal{Y}^*$ is the set of possible productivity levels. A manager observes the worker's *group membership* $G_i \in \{b, w\}$ and a vector of other attributes $S_i^* \in \mathcal{S}^*$ (e.g., educational background, recommendation letters) which we refer to as the *signal*, where $\mathcal{S}^*$ is the set of possible attribute vectors. A worker's attributes and group identity provide information about her productivity. The manager then evaluates the worker by selecting an *action* $A_i^* \in \mathcal{A}^*$ (e.g., hiring decision, offered salary, performance rating), where $\mathcal{A}^* \subset \mathbb{R}$ is the set of possible actions.[10] The manager's payoff depends on her action, the worker's productivity, and (potentially) group identity. She maximizes her expected payoff subject to her beliefs about the joint distribution of productivity, the signal, and group identity. Rather than explicitly modeling the manager's decision problem, we take a reduced-form approach by specifying the manager's action rule $A^* : \mathcal{S}^* \times \{b, w\} \to \mathcal{A}^*$, which determines how the observed signal and group identity maps into an action choice.[11] Given G_i and S_i^*, the manager selects action $A_i^* = A^*(G_i, S_i^*)$.

We embed this labor market decision in a broader economy—a *system*—to capture the idea that a worker's productivity and signal for task n^* may be affected by actions in other domains (e.g., financial, criminal, or past employment) or time periods. The system consists of a set of nodes $\mathcal{N} \equiv \{1, ..., N\} \cup \{n^*\}$, where we refer to n^* as the *focal node*. Each non-focal node $n = 1, ..., N$ corresponds to a task similar in structure to n^*. Specifically, at each n, a

[10] We assume the action is a real number to focus on expected action as a measure of discrimination and we assume the set of groups is binary to simplify notation; the analysis easily extends to more general spaces.

[11] The framework can accommodate a random action rule that induces a distribution over $\mathcal{A}^*$. For simplicity, we focus on a deterministic action rule, as well as abstract from interactions across workers and other realistic features of labor markets.

worker is evaluated for a task for which he has productivity $Y_i^n \in \mathcal{Y}^n$ (e.g., loan repayment probability, criminal activity propensity, past job performance). An evaluator (e.g., loan officer, judge, past manager) observes the worker's group G_i and a signal $S_i^n \in \mathcal{S}^n$, then selects action $A_i^n \in \mathcal{A}^n$ (e.g., loan terms, criminal charges, performance evaluation). As above, the sets $\mathcal{Y}^n$, $\mathcal{S}^n$, and $\mathcal{A}^n$ give the possible productivity levels, attribute vectors, and actions, respectively, and $A^n(G_i, S_i^n)$ denotes the evaluator's action rule.[12] At the end of this subsection we discuss two additional types of decisions this system can include: *(i)* nodes where the decision is impacted by an anticipated action at a future node, and *(ii)* nodes where the decision is the choice of signaling technology for another node.

Actions at one node can impact the productivity and signal at others. For example, the signal S_i^* at the focal node n^* may include a performance evaluation A_i^n from another node n. Similarly, focal-node productivity Y_i^* may be a function of productivity Y_i^n and on-the-job training A_i^n at node n. For simplicity, we maintain that group membership G_i remains the same across all nodes.

We first define *direct discrimination*, which captures group-based differences in focal-node action choices at a given node holding fixed the signal. It occurs when the action rule prescribes different actions for group w and b workers with the same signal realization:

Definition 1 (Direct Discrimination)**.** *Direct discrimination occurs at node $n \in \mathcal{N}$ for signal $s \in \mathcal{S}^n$ if $A^n(w, s) \neq A^n(b, s)$.*[13]

Direct discrimination arises from the worker's group identity itself. It can arise from the manager's preferences, beliefs about productivity, or beliefs about the signal distribution being dependent on group identity (e.g., taste-based or statistical discrimination). Because this definition conditions on the signal at the node where discrimination is being measured, it does not account for disparities that stem from decisions made at other nodes.[14]

We next aim to define a notion of discrimination that incorporates how discrimination at other nodes contributes to disparities at the focal node. A researcher interested in identifying and quantifying such broader inequities must make a key analytic choice: namely, which systemic forces to study. The researcher may wish to hone in on a subset of nodes in order to isolate how actions within this subset contribute to disparities at the focal node—perhaps due to the availability of detailed decision data for these nodes, policy-makers' capacity to

[12] We don't place any restrictions on the sets $\mathcal{Y}^*$, $\mathcal{S}^*$, $\mathcal{Y}^n$, $\mathcal{S}^n$, and $\mathcal{A}^n$ aside from that it is possible to endow each set with a σ-algebra so that probability measures over each set are well-defined.

[13] In a slight abuse of notation, when we write statements across all $n \in \mathcal{N}$, the implied variables for the case of $n = n^*$ are superscripted by n^*, e.g., A^{n^*} corresponds to A^*, etc.

[14] We abstract from some important conceptual issues with considering group membership G_i as separately manipulable in the action rule from S_i^*. Such issues can be especially salient when G_i is meant to capture race; see, e.g., Fryer and Levitt (2004), Sen and Wasow (2016), Gaddis (2017), and Kohler-Hausmann (2019). Notably, the constructivist framework of Rose (2023) can be incorporated to address this issue given a salient and manipulable signal of group membership (e.g., distinctively racial names).

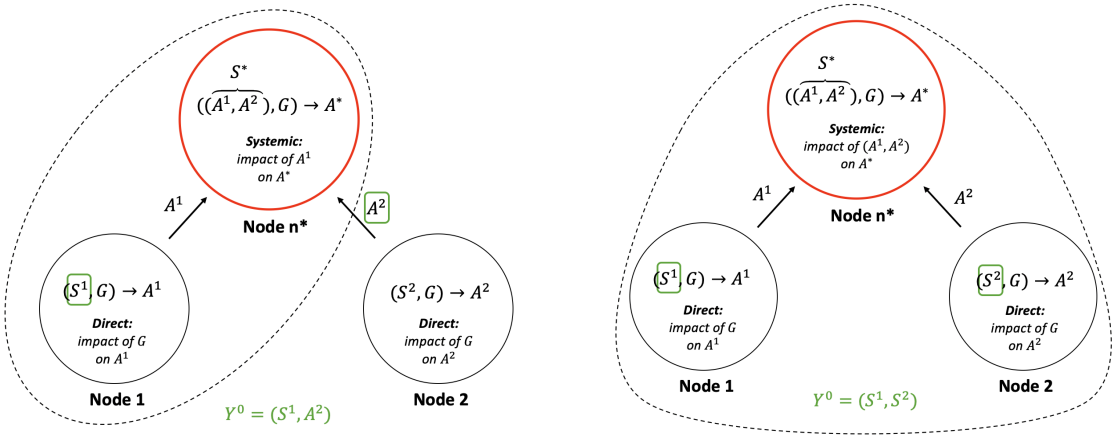
intervene at these nodes, or simply because these are the systemic forces of interest to the researcher. This amounts to choosing a subset of other nodes $\mathcal{N}^0 \subset \mathcal{N}$ to include in the analysis, which we refer to as the *subsystem*, and a measure to group comparable workers at entry into this subsystem, which we refer to as the *reference qualification*. Given a chosen reference qualification, let Y_i^0 denote worker i's qualification level and $\mathcal{Y}^0$ denote the set of possible qualification levels.

Figure 1 illustrates two possible choices of subsystem in a three-node system $\mathcal{N} = \{1, 2, n^*\}$. As in Section 2, suppose the focal node n^* corresponds to a hiring decision for a software company and node 1 corresponds to a platform performance rating (e.g., a GitHub profile). The platform signal S_i^1 corresponds to the programmer's submitted code and the platform action A_i^1 corresponds to her rating. Suppose node 2 corresponds to a software development internship, where the programmer interviewed for the position (signal S_i^2) and either obtained experience or not, $A_i^2 \in \{0, 1\}$. The hiring node signal then consists of the platform rating and the presence or absence of internship experience: $S_i^* = (A_i^1, A_i^2)$. A researcher interested in studying how bias in the platform rating (node 1) contributes to hiring disparities (node n^*) would include these two nodes in the subsystem: $\mathcal{N}^0 = \{1, n^*\}$, as illustrated in Panel (a). Setting the reference qualification as the platform node signal (submitted code) and internship node action (internship experience, since this node is not part of the subsystem) groups comparable workers at entry to this subsystem: $Y_i^0 = (S_i^1, A_i^2)$. This is the implicit choice of Economist 2 in Section 2. If the researcher wants to also study how direct discrimination in internship experience impacts hiring disparities, she would also include node 2 in the subsystem: $\mathcal{N}^0 = \{1, 2, n^*\}$, as illustrated in Panel (b). Now, setting the reference qualification as the platform node signal and internship node *signal*, $Y_i^0 = (S_i^1, S_i^2)$, groups comparable workers at entry to this subsystem. We further discuss the choice of subsystem and qualification in Section 3.3.[15]

Fixing a subsystem and reference qualification, we define a measure of *total discrimination* that captures group-based differences in actions at the focal node for workers who enter the subsystem with comparable qualifications. Worker i enters the subsystem with qualification level Y_i^0 and is evaluated at each node, which potentially impacts his productivity and signal at subsequent nodes. These interactions generate a distribution over how the focal-node productivity Y_i^* and signal S_i^*—and hence action A_i^*—vary by group for workers who start with the same qualification level. Let $\alpha^*(y^0, g) \in \Delta(\mathcal{A}^*)$ denote this action distribution for a group g worker with qualification level y^0, where $\Delta(\mathcal{A}^*)$ denotes the set of distributions over

[15]Note that within a given system, the choice of reference qualification pins down the subsystem. We emphasize both choices as, conceptually, we view it more natural to first choose which systemic forces to focus on (i.e., the subsystem) and then to decide on the measure that ensures comparability at entry to the subsystem (i.e., reference qualification). This also highlights the importance of modeling the system of interactions which give rise to total and systemic discrimination, as defined below, as these structural relationships are crucial for policy evaluation.

FIGURE 1. Two Subsystem Choices in a Three-Node System



(a) Two-node subsystem $\mathcal{N}^0 = \{1, n^*\}$ capturing systemic impact of direct discrimination at node 1

(b) Three-node subsystem $\mathcal{N}^0 = \{1, 2, n^*\}$ capturing systemic impact of direct discrimination at nodes 1 and 2

$\mathcal{A}^*$, and let $\mu^*(y^0, g) \in \mathbb{R}$ denote the corresponding average action.[16] Total discrimination is the difference between the average actions faced by group w and b workers who enter the subsystem with the same qualification level:[17]

Definition 2 (Total Discrimination)**.** *Total discrimination at node n^* for workers with qualification level $y^0 \in \mathcal{Y}^0$ is equal to $\Delta^T(y^0) \equiv \mu^*(y^0, w) - \mu^*(y^0, b)$. Total discrimination arises if $\Delta^T(y^0) \neq 0$ for some $y^0 \in \mathcal{Y}^0$.*

By comparing workers with the same qualification level, total discrimination captures focal node disparities that arise due to decisions within the subsystem. Importantly, we do not claim that the residual focal-node disparity is *not* discrimination; this remaining disparity may stem from decisions outside of the subsystem, non-discriminatory forces, or both. Returning to Figure 1, total discrimination for the subsystem in Panel (a) includes the impact of direct discrimination at node 1 on disparities at node n^*, while in Panel (b) it includes the impact of direct discrimination at both nodes 1 and 2. In Section 3.3 we discuss how, through the choice of subsystem and reference qualification, this definition nests common notions of discrimination in the literature.

Total discrimination includes both the impact of direct discrimination at other nodes and at the focal node.[18] We next define a measure of *systemic discrimination* to isolate

[16] Formally, $\alpha^* : \mathcal{Y}^0 \times \{b, w\} \to \Delta(\mathcal{A}^*)$ and $\mu^* : \mathcal{Y}^0 \times \{b, w\} \to \mathbb{R}$ are mappings from the sets of qualification levels and group memberships to a distribution over actions or an expected action, respectively.

[17] Our definitions based on means easily generalize to other distributional features of α^*, such as variances or higher-order moments. We focus on means for notational simplicity and since these yield the most commonly studied discrimination measures in practice.

[18] That total discrimination includes the impact of focal-node direct discrimination can be seen from (1): the dependence of α^* on A^* means that direct discrimination at n^*—as captured by $A^*(b, s) \neq A^*(w, s)$—can

the component of total discrimination that arises indirectly from discrimination at other nodes. In particular, this measure constructs a counterfactual action distribution to compare average actions when both groups face the same focal-node action rule—thereby shutting down the possibility of direct discrimination. Let $\sigma^*(y^0, g) \in \Delta(\mathcal{S}^*)$ denote the node-n^* signal distribution for a group g worker with qualification level y^0. Note that the action distribution $\alpha^*(y^0, g)$ depends on $\sigma^*(y^0, g)$ and the action rule $A^*(g, s)$ for group g:

$$\alpha^*(A; y^0, g) = \sigma^*(\{s : A^*(g, s) \in A\}; y^0, g) \tag{1}$$

for any measurable set $A \subset \mathcal{A}^*$.[19] Define the counterfactual action distribution $\tilde{\alpha}^*(y^0, g) \in \Delta(\mathcal{A}^*)$ for a group g worker with qualification level y^0 as the action distribution under signal distribution $\sigma^*(y^0, g)$ and action rule $A^*(g', s)$ for group $g' \neq g$:

$$\tilde{\alpha}^*(A; y^0, g) \equiv \sigma^*(\{s : A^*(g', s) \in A\}; y^0, g) \tag{2}$$

for any measurable set $A \subset \mathcal{A}^*$. Let $\tilde{\mu}^*(y^0, g)$ denote the corresponding average action. Systemic discrimination compares the average action $\mu^*(y^0, w)$ of group w workers to the counterfactual average action $\tilde{\mu}^*(y^0, b)$ of group b workers under the group w action rule—in other words, the average action for group b versus w when both groups face the group w action rule (or vice versa)—holding fixed the qualification level:

Definition 3 (Systemic Discrimination)**.** *Systemic discrimination at node n^* for workers with qualification level y^0 is equal to $\Delta^S(y^0, w) \equiv \mu^*(y^0, w) - \tilde{\mu}^*(y^0, b)$ under the group w action rule or $\Delta^S(y^0, b) \equiv \tilde{\mu}^*(y^0, w) - \mu^*(y^0, b)$ under the group b action rule. Systemic discrimination arises at n^* if $\Delta^S(y^0, w) \neq 0$ or $\Delta^S(y^0, b) \neq 0$ for some $y^0 \in \mathcal{Y}^0$.*

Varying the choice of subsystem and reference qualification provides measures of systemic discrimination that capture the impact of different sets of systemic forces. While systemic and total discrimination are defined for a particular qualification level, it is also possible to construct an overall measure by averaging across qualification levels.[20]

To illustrate the definitions, consider the analysis of Economist 2 in Section 2. As discussed above, she selects a two-node subsystem with the hiring node (focal node n^*) and the platform node, $\mathcal{N}^0 = \{1, n^*\}$, and she sets the platform signal (i.e., code submitted to the platform) as the reference qualification: $Y_i^0 = S_i^1$. Suppose the platform's performance pro-

translate to different action distributions.

[19] Given a distribution f with support $\mathcal{X}$ parameterized by (y^0, g), we write $f(y^0, g)$ to denote the distribution over $\mathcal{X}$, where $f(y^0, g) \in \Delta(\mathcal{X})$, and $f(X; y^0, g)$ to denote the probability mass $f(y^0, g)$ assigns to measurable set $X \subset \mathcal{X}$, where $f(X; y^0, g) \in [0, 1]$.

[20] The interpretation of this overall measure depends on the qualification distribution used to take the average: using the population qualification distribution measures average discrimination across both groups, while using the qualification distribution for group g measures average discrimination for group g workers.

file is a rating of 0 to 100 (e.g., a code acceptance rate) and that gender bias on the platform leads female programmers to receive systematically lower ratings than male programmers for similar code. Suppose further that the company hires male programmers with a rating above 90 and female programmers with a rating above 80. Thus, there is direct discrimination at both the platform node (from the biased rating, favoring male programmers) and the hiring node (from the differential hiring thresholds, favoring female programmers).

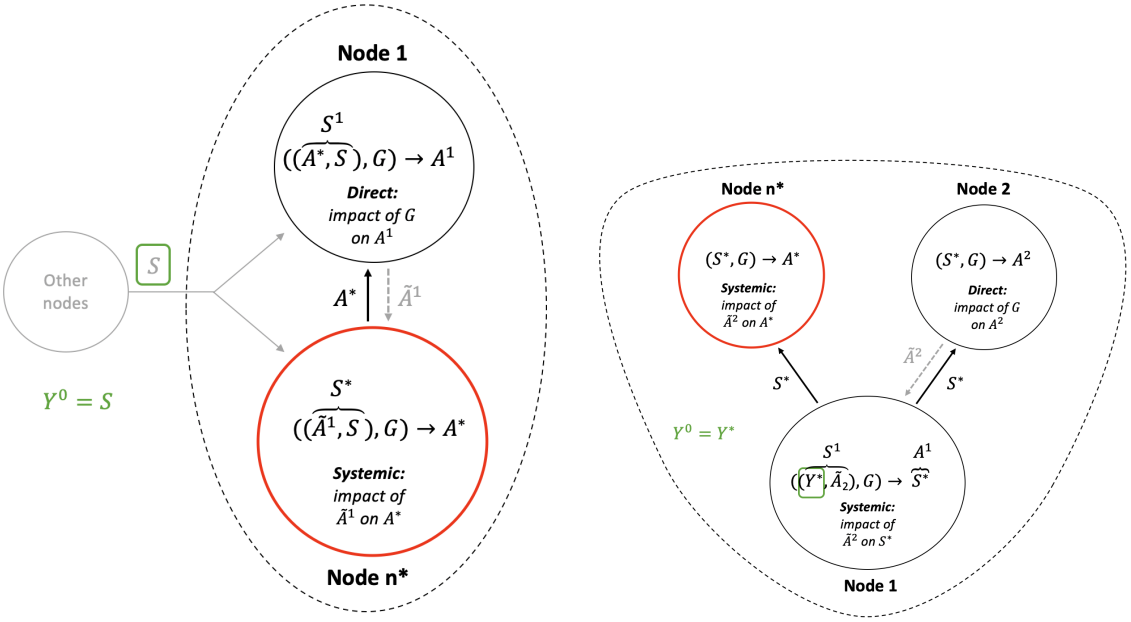
Systemic discrimination captures hiring disparities due to the direct discrimination at the platform node: fixing the hiring threshold as either that used for males (90) or females (80), it compares the probability that females and males who submit similar code have a rating above this threshold. Because of the direct discrimination in ratings, this hiring probability is higher for male programmers than female programmers. Total discrimination combines this with the direct discrimination in hiring: it compares the probability that a male programmer has a rating above 90 to the probability that a female with similar code has a rating above 80. Economist 2 finds that this hiring probability is higher for male programmers than female programmers, because the systematic bias in platform ratings outweighs the 'reverse' direct discrimination in hiring thresholds.

Additional Types of Decision Nodes. Before proceeding, we make two notes on the generality of this framework. First, while the examples so far have focused on systemic discrimination from past actions, our framework can also capture systemic discrimination driven by how anticipated actions at contemporaneous or future nodes impact the focal-node decision. For example, the expected ruling of a jury may impact a defense attorney's decision to accept a plea deal. This can be captured by including the anticipated action $\tilde{A}_i^n$ as part of the focal-node signal S_i^*. This anticipated action may be correct, as in a Nash equilibrium, or incorrect, as in alternative solution concepts (e.g., self-confirming or Berk-Nash equilibrium). Figure 2(a) depicts a subsystem with this type of node. The decision at node 1 is made subsequent to the decision at node n^*, and the anticipated action $\tilde{A}_i^1$ is relevant for A_i^*. The reference qualification (in green) is the component of the node-n^* signal generated outside of the subsystem (i.e., S_i^* excluding $\tilde{A}_i^1$). Systemic discrimination captures the impact of expected direct discrimination at node 1 via $\tilde{A}_i^1$ on A_i^*.[21]

Second, while we have so far focused on settings where systemic discrimination arises from differences in the distribution of a given set of signals (i.e., for a fixed evaluation criteria), the framework can also capture systemic discrimination arising from the choice of the productivity signaling technology itself (i.e., choosing the evaluation criteria). This can be modeled as a decision node 1, where the signal corresponds to focal-node productivity,

[21] For example, Avery et al. (2023) find that the use of artificial intelligence (AI) in recruitment increases the share of women applicants, as women anticipate less bias when assessed by AI instead of humans. Ruebeck (2024) shows that increasing perceived discrimination by revealing that managers mostly promoted White men lowered the retention of female workers and increased their reservation wages.

FIGURE 2. Other Forms of Systemic Discrimination



(a) Two-node subsystem capturing the systemic impact of predicted direct discrimination at future node 1

(b) Three-node subsystem capturing the systemic impact of predicted direct discrimination at node 2 via evaluation design node 1

$S_i^1 = Y_i^*$, the action corresponds to the focal-node signal, $A_i^1 = S_i^*$, and the (random) action rule maps productivity and group to a distribution $\Delta(\mathcal{S}^*)$ over focal-node signals.[22] Figure 2(b) depicts a subsystem with this type of node. At node 1, the worker chooses a signaling technology (e.g., reported traits on a job placement questionnaire); this signal impacts actions at nodes n^* and 2 (e.g., a job hiring decision and a social evaluation). The anticipated action $\tilde{A}_i^2$ impacts the chosen signaling technology at node 1, and hence, the node n^* action A_i^*. The reference qualification is productivity, $Y_i^0 = Y_i^*$, which is the node-1 signal (e.g., actual traits). Systemic discrimination captures the impact of expected direct discrimination at node 2 on the choice of signaling technology, and hence, the focal-node hiring decision A_i^*.[23]

[22] De Plevitz (2007) and Pincus (1996) discuss this design choice as a channel for systemic discrimination, where an institution is first "designed" by a group in power leading to the development of evaluation criteria that are optimized around the characteristics of this group. For example, by not accounting for the family structure and cultural obligations of the Aboriginal community, De Plevitz (2007) discusses how the "Euro-centric model of teaching" creates systemic barriers in schooling for Aboriginal children. Another example is the practice of excluding women from medical trials, which leads to a less informative signal of the efficacy of new treatments for these groups (Bierer et al. 2022).

[23] This example takes inspiration from Bursztyn et al. (2017), where single women report lower desired salaries and willingness to work long hours on a job placement questionnaire (node 1) when the answers were observable to their classmates, as such traits are viewed as undesirable for women (but not men) on the marriage market (node 2). This then led to disparities in job market evaluations (node n^*), where such traits are viewed as desirable for both groups.

3.2 Decomposing Total Discrimination

The three forms of discrimination are linked by decompositions of total discrimination into direct and systemic components. From Definitions 2 and 3, $\Delta^T(y^0)$ measures total discrimination and $\Delta^S(y^0, w)$ or $\Delta^S(y^0, b)$ measure systemic discrimination at qualification level y^0. From Definition 1, $\delta(s) \equiv A^*(w, s) - A^*(b, s)$ measures direct discrimination at n^* following signal realization s. A measure of *average direct discrimination* for workers with the same qualification level is thus given by the expectation of direct discrimination with respect to the signal distribution for this qualification level:

$$\Delta^D(y^0, g) \equiv \int_{\mathcal{S}} \delta(s) d\sigma^*(s; y^0, g) \tag{3}$$

for $g \in \{w, b\}$, where $\sigma^*(y^0, g)$ is the group-specific signal distribution defined above. Similar to how the systemic discrimination measures depend on group via which action rule is fixed, average direct discrimination depends on group via which signal distribution is fixed.

Given these measures, we have the following result:

Proposition 1. *At each qualification level y^0, total discrimination is the sum of average direct discrimination and systemic discrimination:*

$$\underbrace{\Delta^T(y^0)}_{\textit{Total discrimination}} = \underbrace{\Delta^D(y^0, w)}_{\textit{Avg. direct discrimination}} + \underbrace{\Delta^S(y^0, b).}_{\textit{Systemic discrimination}} \tag{4}$$

Equation (4) follows by adding and subtracting $\tilde{\mu}^*(y^0, w)$ to the definition of $\Delta^T(y^0)$ and rearranging terms.[24] It is in the spirit of Kitagawa (1955), Oaxaca (1973), and Blinder (1973), who relate unconditional disparities to a component explained by observable worker characteristics (e.g., education or labor market experience) and an "unexplained" disparity. These classic decompositions can be viewed as a strategy for measuring direct discrimination, while Equation (4) instead leads to strategies for measuring systemic discrimination and relating it to direct discrimination (see Section 4).

As with the classic Kitagawa-Oaxaca-Blinder approach, there are multiple ways to decompose total discrimination into direct and systemic components. We can also express total discrimination as the sum of average direct discrimination with respect to the group b signal distribution and systemic discrimination under the group w action rule:[25]

$$\Delta^T(y^0) = \Delta^D(y^0, b) + \Delta^S(y^0, w). \tag{5}$$

[24] Note that $\mu^*(y^0, w) = \int_{\mathcal{A}^*} a d\alpha^*(a; y^0, w) = \int_{\mathcal{S}^*} A(w, s) d\sigma^*(s; y^0, w)$ and $\tilde{\mu}^*(y^0, w) = \int_{\mathcal{A}^*} a d\tilde{\alpha}^*(a; y^0, w) = \int_{\mathcal{S}^*} A(b, s) d\sigma^*(s; y^0, w)$. Then $\Delta^T(y^0) = \mu^*(y^0, w) - \mu^*(y^0, b) + \tilde{\mu}^*(y^0, w) - \tilde{\mu}^*(y^0, w) = \int_{\mathcal{S}^*} A(w, s) d\sigma^*(s; y^0, w) - \int_{\mathcal{S}^*} A(b, s) d\sigma^*(s; y^0, w) + \Delta^S(y^0, b) = \Delta^D(y^0, w) + \Delta^S(y^0, b)$.

[25] This follows by adding and subtracting $\tilde{\mu}^*(y^0, b) = \int_{\mathcal{S}} A(w, s) \sigma^*(s; y^0, b) ds$ to and from the definition of $\Delta(y^0)$ and rearranging terms.

Averaging Equations (4) and (5) yields a third decomposition:

$$\Delta^T(y^0) = \overline{\Delta}^D(y^0) + \overline{\Delta}^S(y^0), \tag{6}$$

where $\overline{\Delta}^D(y^0) \equiv \frac{1}{2}(\Delta^D(y^0, w) + \Delta^D(y^0, b))$ and $\overline{\Delta}^S(y^0) \equiv \frac{1}{2}(\Delta^S(y^0, w) + \Delta^S(y^0, b))$.

3.3 Discussion

Systemic Discrimination in Other Fields. Our formalization of systemic discrimination aligns broadly with indirect forms of discrimination discussed in the sociology literature: i.e., forms of inequality operating indirectly through characteristics beyond group identity and stemming from discrimination in other parts of a system. As discussed in Small and Pager (2020), part of the challenge of assessing different notions of indirect discrimination in prior work is that different terms—such as "structural," "systemic," and "institutional" discrimination—are used to describe this general phenomenon. Pincus (1996), for example, defines structural discrimination as policies "which are race/ethnic/gender neutral in intent but which have a differential and/or harmful effect on minority race/ethnic/gender groups" (see also Hill 1988).[26] Similarly, Powell (2008) defines systemic discrimination as a "product of reciprocal and mutual interactions within and between institutions," both "within and across domains."[27] Our formalization captures both of these definitions: we measure systemic discrimination as disparities under an action rule $A^*(g, \cdot)$ that is fixed across groups, which arise due to discriminatory decisions at other nodes.[28] Our measure also captures both types of indirect discrimination in Feagin (1977)'s typography: past-in-present discrimination, which arises indirectly from discriminatory actions at past nodes (as in Figure 1) and side-effect discrimination, which arises indirectly from discriminatory actions in other domains.[29] Appendix A reviews other connections to literatures on systemic discrimination.

Choosing a Subsystem and Reference Qualification. The interpretation of the systemic and total discrimination measures is tied to the researcher's chosen subsystem and reference qualification. Through these choices, the measures can bring focus to different systemic forces and study different forms of discrimination.

Three examples illustrate this feature of the framework. First, by including only the

[26] For example, the historical practice of "redlining" in mortgage markets prioritized borrowers from majority-White neighborhoods over equally-creditworthy borrowers from majority-Black neighborhoods. Such neighborhood-based prioritization generated substantial race-based lending disparities despite the policy typically being prima facie race-neutral. In our framework, any total discrimination that arises from a race-neutral action rule is systemic, as the direct component is zero by definition.

[27] He terms discrimination arising from the interactions of systems as "structural" and discrimination stemming from interactions in a system as "systemic." We do not formalize this distinction here.

[28] Our definition also aligns with some notions of institutional discrimination (Small and Pager 2020).

[29] Notably, while Feagin (1977) referred to both forms of discrimination as indirect institutional discrimination, Feagin (2013) refers to both as forms of systemic discrimination.

focal node in the subsystem ($\mathcal{N}^0 = \{n^*\}$) and choosing the focal-node signal as the reference qualification ($Y_i^0 = S_i^*$), total discrimination is aligned with direct discrimination; it does not account for any systemic forces. This is the choice of Economist 1 in Section 2 and corresponds to model (I) in Cain (1986)'s classification of statistical models for measuring discrimination. Second, choosing focal-node productivity as the reference qualification ($Y_i^0 = Y_i^*$) corresponds to including all nodes that impact how workers signal this productivity at node n^*. This aligns total discrimination with what Arnold et al. (2022) argue captures the legal notion of *disparate impact*.[30] It also relates to some measures of algorithmic unfairness, e.g., Berk et al. (2021).[31] This is the choice of Economist 3 in Section 2, who sets Y_i^0 to be coding ability. Finally, including all nodes in the subsystem ($\mathcal{N}^0 = \mathcal{N}$) and setting the reference qualification to a constant presumes that there are no differences prior to entry in the system and all differences emerge from systemic forces. Total discrimination is then the unconditional disparity between groups and the systemic component accounts for the impact of all actions influencing the focal-node productivity and signal. This choice corresponds to model (II) in the taxonomy of Cain (1986).

Other choices can isolate different systemic forces in the economy. To measure the impact of specific past discriminatory decisions on treatment at the focal node (i.e., past-in-present discrimination) one can include the relevant node(s) in the subsystem and set the reference qualification to be information at entry to the subsystem, as illustrated in Figure 1. This is the choice of Economist 2 in Section 2, who sets Y_i^0 to be the information available to the online platform. Similarly, one can include decision nodes from other domains to study how discriminatory practices in one domain impact treatment in another (i.e., side-effect discrimination). Finally, one can include nodes where decisions are made contemporaneous with or subsequent to the focal node, where anticipated discrimination at these nodes impacts action choices at the focal node, as illustrated in Figure 2.[32]

Thus, through the choice of subsystem and qualification, Definitions 1 to 3 provide a common framework for studying different notions of discrimination in the literature and—as

[30] The legal notion of disparate impact is captured in the landmark Supreme Court decision *Griggs v. Duke Power Co.* (1971), which found that Duke Power's policy of requiring a high school diploma for within-company transfers was discriminatory because it disadvantaged Black employees who were qualified but lacked a degree, in part due to ongoing discrimination in secondary education. The Court noted that the degree requirement bore no relevance to an individual's ability to perform different jobs at the company. Notably, discrimination was found despite the policy being facially "race-blind": White and Black employees with the same educational background had the same ability to transfer jobs.

[31] Strack and Yang (forthcoming) characterize the set of signals that satisfy a 'privacy-preserving' property. Applying this characterization to our framework yields the set of signals that do not generate disparate impact for any action rule.

[32] An open debate is whether group-based differences in preferences that generate productivity or signal differences should be coded as discrimination. For example, gender socialization may lead female workers to refrain from asking for a raise, which then impacts their benchmark for future wage setting (e.g., Cook et al. 2021). The subsystem and qualification can be chosen to include or exclude such preference differences.

in the motivating example—can help interpret and integrate seemingly disparate findings.

Sources of Systemic Discrimination. Systemic discrimination can stem from an *informational* source, via group-based differences in the focal-node signal for workers who are equally productive at the focal node, or a *technological* source, via group-based differences in focal-node productivity for workers who are equally qualified at entry to the subsystem. The former captures the impact of other decisions on the ability to signal productivity, while the latter captures the impact of other decisions on the ability to build productivity.

Informational systemic discrimination can stem from *signal inflation*, where the signal is, on average, higher for one group than the other, conditional on focal node productivity. An example is when bail decisions depend on criminal records, and Black defendants have longer criminal records than White defendants with the same pretrial misconduct potential due to direct discrimination in other law enforcement interactions.[33] Other examples include recruitment practices that prioritize workers with certain social connections, where one group is more connected than equally qualified members of the other, and wage setting based on salary history, where one group has higher past salaries than equally productive members of the other (Agan et al. 2024). It can also arise from *differential screening*, where the manager has a more precise signal for one group than the other (Cornell and Welch 1996).[34] For example, a hiring process optimized for men may generate less informative signals for women (Pinkston 2003; Mocanu 2023).[35] Another example is how discrimination in access to past borrowing opportunities leads to differentially informative credit histories amongst borrowers with the same ability to repay (Bartik and Nelson 2024). See Appendix B for an empirical illustration of differential screening.

Similarly, technological systemic discrimination emerges when differences in productivity arise from discrimination at other nodes. For example, White workers might have more opportunities to build human capital than Black workers due to discrimination in access to training and skill development.[36] Alternatively, Black workers may respond to anticipated future direct discrimination by investing less in human capital (Lundberg and Startz 1983; Coate and Loury 1993).[37] Technological systemic discrimination also includes the type of

[33]For example, Pierson et al. (2020) show that Black individuals are more likely to be stopped by police and charged with a crime. Such criminal record disparities have also been shown to contribute to systemic discrimination in the labor market (Pager et al. 2009; Agan and Starr 2017).

[34]In Cornell and Welch (1996), minority job applicants receive fewer signal draws than majority applicants; there is no direct discrimination since applicants with the same number of draws are treated equally.

[35]Similarly, De Plevitz (2007) documents racial disparities in job screening stemming from using height-to-weight ratios calibrated with Anglo-Celtic data.

[36]For example, Gallen and Wasserman (2021) document gender differences in career advice, where women are more likely to receive advice about work/life balance than men. This can deter investment in human capital and the pursuit of careers in competitive fields.

[37]Here, workers have no group-based differences in initial productivity (Y_i^0). They make a costly decision to invest in human capital that increases productivity at the focal node (Y_i^*).

"task-based" discrimination studied in Hurst et al. (2024), where racial barriers to task specialization generate group-based differences in choice of specialization. Importantly, our framework allows researchers to develop designs for bringing these concepts to data.[38]

Understanding the source of systemic discrimination is crucial for identifying an appropriate way to counteract it. For example, since signal inflation is a statistical bias in the signal, it can be offset via "reverse" direct discrimination in the action rule. In the case discussed earlier, if the bail judge is aware of how direct discrimination contributes to criminal record disparities, then she can account for it in her interpretation of criminal records. Hence, one potential policy response is ensuring awareness of signal inflation. In contrast, differential screening and technological systemic discrimination cannot be offset by simply adjusting the action rule without payoff consequences. To offset differential screening, more information needs to be collected for one group or the additional information needs to be ignored for the other group, creating a trade-off between equity and accuracy. Similarly, offsetting technological systemic discrimination requires additional investment in training for the disadvantaged group.

Relation to Statistical Discrimination. Our measure of systemic discrimination stems from group-based differences in the signal and productivity distributions, as with classical statistical (direct) discrimination. Statistical discrimination differs from systemic discrimination, however: it arises from the impact of these distributions on the action rule for a given signal realization, rather than the action distribution for a given qualification level. When $Y_i^0 \neq S_i^*$, differences in the signal distribution can lead to both systemic and statistical discrimination, and similarly for differences in the productivity distribution when $Y_i^0 \neq Y_i^*$ or $Y_i^0 \neq S_i^*$. In both cases, focusing only on direct discrimination can miss a key aspect of how group differences in these distributions contribute to action disparities.[39]

4 Identification

We now show how the framework can be brought to data with the iterated audit (IA) approach. We formalize the approach with a two-node subsystem with the reference qualifi-

[38]Consider, for example, a two-node system with a hiring focal node and a node corresponding to the acquirement of job-relevant human capital. A researcher studying this system with a constructive IA design (discussed below) could learn whether racial disparities in hiring are due to a *choice* made by workers to invest in human capital in anticipation of future discrimination (as in Coate and Loury 1993 and Figure 2(a)), or whether hiring disparities are due to direct discrimination in the opportunity to obtain human capital (as in Hurst et al. 2024 and Figure 1(a)).

[39]Unlike with statistical discrimination (e.g., Bordalo et al. 2019; Bohren et al. 2023) there is no scope for "inaccurate" systemic discrimination: only the true productivity and signal distributions contribute to systemic discrimination. However, inaccurate beliefs about these distributions can lead to inaccurate perceptions about the extent of systemic discrimination. This can affect action choices, and hence, total discrimination. For example, a mortgage assessor may incorrectly believe that a credit score is an unbiased signal of creditworthiness for groups, and therefore fail to adjust for its bias.

cation chosen to be the initial-node signal, $Y_i^0 = S_i^1$. We define $S^*(a^1, y^0)$ as the focal-node signal that is realized given an initial-node action of a^1 among those with $Y_i^0 = y^0$, such that $S_i^* = S^*(A_i^1, S_i^1)$. We assume that the researcher observes Y_i^0, A_i^1 and S_i^*.[40] We discuss at the end of this section how the approach extends to multi-node subsystems and other choices of reference qualification.

4.1 Treatment and Interaction Components

Measuring total and systemic discrimination involves the measurement of two conceptually distinct components: *(i)* a treatment component capturing direct discrimination at focal and non-focal nodes and *(ii)* an interaction component capturing how the actions at non-focal nodes impact the action at the focal node via focal-node signals. Direct discrimination, and hence the treatment component, can be identified by conventional experimental designs such as a standard audit or correspondence study. The key challenge to measuring total and systemic discrimination is therefore identification of the interaction component.

To formalize the identification challenge, recall that total discrimination is measured by $\Delta^T(y^0) = \mu^*(y^0, w) - \mu^*(y^0, b)$ and systemic discrimination is measured by $\Delta^S(y_0, w) = \mu^*(y^0, w) - \tilde{\mu}^*(y^0, b)$ or $\Delta^S(y_0, b) = \tilde{\mu}^*(y^0, w) - \mu^*(y^0, b)$. In our two-node system with $Y_i^0 = S_i^1$ and $S_i^* = S^*(A_i^1, S_i^1)$, the objects in these measures can be written as:

$$\begin{aligned} \mu^*(y^0, g) &= E[A^*(g, S_i^*) \mid Y_i^0 = y^0, G_i = g] \\ &= A^*(g, S^*(A^1(g, y^0), y^0)) \qquad (7) \\ \tilde{\mu}^*(y^0, g) &= E[A^*(g', S_i^*) \mid Y_i^0 = y^0 G_i = g] \\ &= A^*(g', S^*(A^1(g, y^0), y^0)), \qquad (8) \end{aligned}$$

recalling that $A^*(g, s)$ and $A^1(g, s)$ are the action rules at the focal and initial node, respectively. Clearly, knowing how group membership directly affects actions at the focal and initial nodes (formally, how $A^*(g, s)$ and $A^1(g, s)$ depend on g) is not enough to identify equations (7) and (8). It is also necessary to know how initial-node actions indirectly affect focal-node actions through focal-node signals (i.e., how $A^*(g, S^*(a^1, y^0))$ depends on a^1). The former is what we term knowledge of the treatment component; the latter is what we term knowledge of the interaction component.

To make Equations (7) and (8) concrete, consider a two-node labor market setting in which workers first apply for an internship (the initial node) and then apply for an entry-level

[40] For ease of notation, we implicitly assume that S_i^1 is the only difference across workers at entry to the subsystem and that the focal-node signal only depends on S_i^1 and A_i^1. In practice, workers could also differ along other components of the focal-node signal S_i^* that are independent of A_i^1, in which case the reference qualification would also include these components. If the researcher did not observe these other components, then ensuring that their distributions are similar across groups (as in an audit/correspondence study) identifies our three measures of discrimination.

position (the focal node). Equations (7) and (8) show that total and systemic discrimination in entry-level hiring actions arises from two distinct components: a treatment component capturing direct discrimination in internship and entry-level hiring and an interaction component capturing how internship experience impacts entry-level hiring. To estimate the former treatment component, a researcher could run a conventional audit study where workers of different races or genders apply to internships with identical resumes. But this audit would not identify the interaction component and hence not identify systemic discrimination.

We develop two iterated audit approaches to estimate the interaction component: a *constructive* approach, which separately estimates the impact of each possible action at the other node on the focal node action, and an *experimental* approach, which directly measures the interaction component by simulating the impact of other-node action distributions.

4.2 The Constructive IA Approach

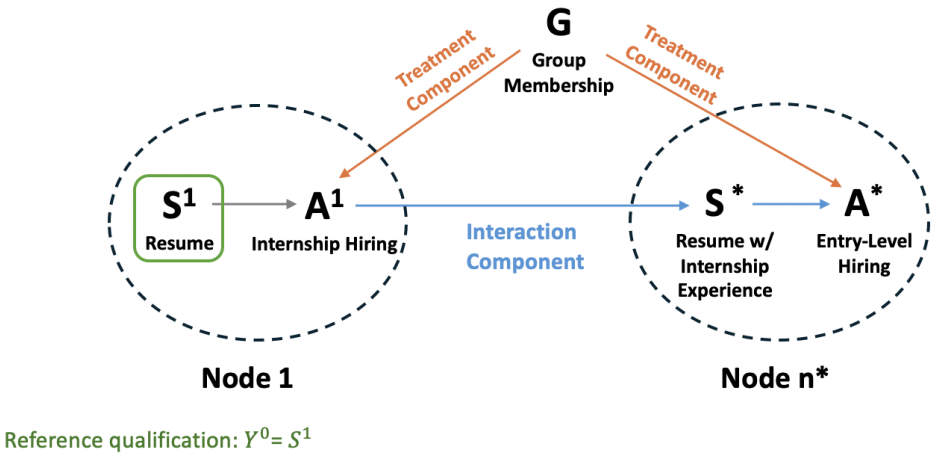
The first IA approach separately estimates the effect of each possible action at non-focal nodes on the focal-node signal, along with the effect of each possible focal-node signal realization on the focal-node action. In our two-node example, this means learning how $S^*(a^1, y^0)$ depends on a^1 and how $A^*(g, s)$ depends on s. With these, a researcher can *construct* an estimate of the interaction component: e.g. how $A^*(g, S^*(a^1, y^0))$ depends on a^1.

To make this constructive IA approach concrete, consider again the internship and entry-level hiring example. A researcher could first learn how past internship experience a^1 translates to the signal $S^*(a^1, y^0)$ that a hiring manager observes when evaluating candidates for the entry-level job; e.g., she could see that candidates with internship experience list it on resumes sent to hiring managers. The researcher could then learn how hiring manager actions depend on this experience; e.g., she could randomize different internship experiences to resumes and measure hiring decisions. Combining these steps, the researcher can construct a measure of how internship experience impacts entry-level hiring. Combining this with conventional audit information on direct discrimination in both internship and entry-level hiring, she could then measure systemic and total discrimination in entry-level hiring.

Figure 3 illustrates the constructive IA design for this example. The treatment component is given by estimates of how group membership directly affects internship and entry-level hiring (the orange arrows). The interaction component is given by estimates of how internship hiring impacts entry-level hiring through resume internship experience (the blue arrows). The combination of these estimated effects identifies total discrimination in entry-level hiring. Removing the impact of the orange arrow on the right, which captures direct discrimination in entry-level hiring, identifies systemic discrimination.

This example also highlights the importance of measuring the interaction component and why conventional audit study estimates of the treatment component are not enough to identify systemic or total discrimination. If the level of internship experience is critical

FIGURE 3. Constructive IA Design in Entry Level Hiring



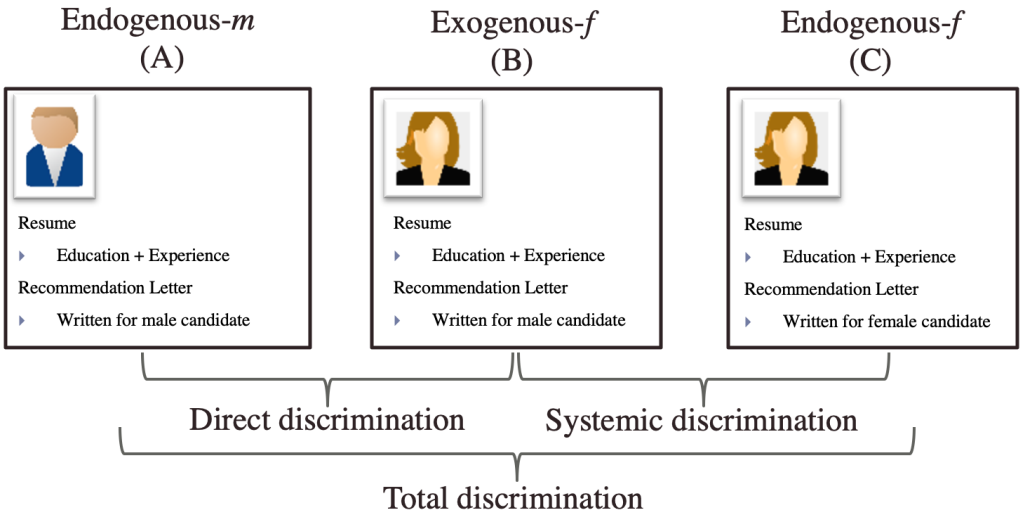
for future success in hiring, then minimal direct discrimination in internship hiring could lead to large systemic and total discrimination in entry-level hiring through the interaction component. More generally, nonlinearities in how actions at non-focal nodes affect focal-node actions means that conventional audit study analyses of direct discrimination at multiple nodes will fail to reveal the total extent of discrimination arising from interactions across nodes. Conversely, decision-makers might act to undo perceived direct discrimination in the signals they observe, such that direct discrimination at non-focal nodes does not translate to total discrimination at the focal node. For example, entry-level hiring managers may put less weight on internship experience for groups that have historically faced discrimination in such opportunities. In this case, even if a conventional audit reveals sizable direct discrimination in internship hiring, it may translate to minimal total discrimination in entry-level hiring.

While conceptually straightforward, the constructive IA approach may be challenging to implement when action or signal spaces are high-dimensional or otherwise complex. Suppose, for example, that workers apply to the entry-level position with both a resume and recommendation letter written by the internship supervisor. This supervisor discriminates in the language and tone of the letters for equally qualified workers. Here the set of signals (i.e., possible recommendation letters) is high-dimensional, making it hard to estimate how $A^*(g, s)$ depends on s. To address such issues, we turn to the second IA approach.

4.3 The Experimental IA Approach

The second IA approach directly measures systemic and total discrimination by simulating the distribution of focal-node signals as impacted by group membership through the actions at other nodes, and measuring focal-node actions given the simulated signal distributions. In our two-node example, measuring $\Delta^S(y^0, w)$ and $\Delta^T(y^0)$ means generating draws of $\tilde{S}_i^* \mid G_i = g, Y_i^0 = y^0$ for $g \in \{w, b\}$ and measuring $A^*(w, \tilde{S}_i^*(w, y^0))$, $A^*(b, \tilde{S}_i^*(b, y^0))$ and

FIGURE 4. Experimental IA Design in Entry-Level Hiring



$A^*(w, \tilde{S}_i^*(b, y^0))$ to *experimentally* measure $\mu^*(y^0, w)$, $\mu^*(y^0, b)$ and $\tilde{\mu}^*(y^0, b)$.[41]

To make this approach concrete, consider a modified version of the previous example where entry-level hiring managers observe recommendation letters from the internship supervisors. A researcher could obtain a set of recommendation letters from individuals of each group. She could then generate entry-level job applications by drawing letters from these sets and attaching a salient signal of group membership—both the same group as the individual that the letter was written for, g, and the other group, g'. The average entry-level hiring manager action for these simulated applications identifies $\mu^*(y^0, g)$ when the application and letter were both assigned group g, and $\tilde{\mu}^*(y^0, g)$ when the application was assigned to group g' but the letter was written for an individual from group g. Together, these estimates identify systemic and total discrimination.

Figure 4 illustrates the experimental IA design for this example, focusing on gender discrimination. Hiring managers are presented with applications from one of three groups which allow the researcher to decompose total discrimination into average direct discrimination with respect to the male signal distribution and systemic discrimination under the female action rule. The first two groups, termed the *endogenous-m* and *exogenous-f* groups, are similar to those used in a standard correspondence or audit study: the resumes and recommendation letters are identical except for a signal of gender (e.g., a distinctively male or female name). Here, the recommendation letters are drawn from the set of male letters. The third group of applicants, termed the *endogenous-f* group, has resumes with distinctively female names but recommendation letters drawn from the set of female letters. Comparing the hiring actions of endogenous-m vs. exogenous-f applications identifies average direct dis-

[41] Analogously, using these simulated distributions to measure $A^*(w, \tilde{S}_i^*(w, y^0))$, $A^*(w, \tilde{S}_i^*(w, y^0))$ and $A^*(w, \tilde{S}_i^*(b, y^0))$ identifies $\Delta^S(y^0, b)$ and $\Delta^T(y^0)$.

crimination since it compares identical resumes and letters, as in a standard correspondence study, while comparing the actions of exogenous-f vs. endogenous-f applications identifies systemic discrimination. Finally, comparing endogenous-m vs. endogenous-f applications identifies total discrimination.

While being more practical in settings with high-dimensional or otherwise complex signals, this approach has the drawback of not separately measuring the impact of initial-node actions on focal-node signals as in the constructive IA approach. Consequently, the precise mechanisms behind systemic discrimination may be harder to infer—which can be important for forming appropriate policy responses. Still, by capturing the interaction component, the experimental IA approach will also account for any nonlinearities or offsetting behavior that might obscure the impact of systemic discrimination if one were to try to infer it from accumulated estimates of direct discrimination at multiple nodes.

4.4 Extensions

It is straightforward to extend both IA approaches to subsystems with multiple nodes and to other observed reference qualifications. The treatment component would again be estimated by the direct effects of group membership on actions at the focal node and all other nodes. In the constructive IA approach, the researcher would learn how the focal-node signal $S^*(a^1, ..., a^N, y^0)$ depends on all other-node actions $(a^1, \ldots, a^N)$, as well as how the focal-node action rule $A^*(g, s)$ depends on s. These estimates can be used to construct the interaction component, i.e., the dependence of $A^*(g, S^*(a^1, ..., a^N, y^0))$ on $(a^1, ..., a^N)$, via generalized versions of Equations (7) and (8). In the experimental IA approach, the researcher would simulate the conditional-on-Y_i^0 distribution of focal-node signals, as impacted by group membership through the actions at all other nodes, and again measure focal-node actions given the simulated signal distributions.

In subsystems with many nodes, an experimental IA may be more tractable than an constructive IA as it avoids the need to estimate the effect of each possible action $a^n \in \mathcal{A}^n$ for $n = 1, ..., N$ on the focal-node signal. On the other hand, the constructive IA approach may be more useful for studying mechanisms, identifying the nodes that generate the biggest systemic disparities, or forming appropriate policy responses, as it separates out each indirect effect that gives rise to systemic discrimination. These tradeoffs mirror those from the discussion of complex or high-dimensional signals in the two-node case.

Extensions to partially observed or unobserved Y_i^0 are likely more involved. Suppose, for example, the reference qualification is chosen to be focal-node productivity: $Y_i^0 = Y_i^*$. In some settings, productivity may be observed only selectively: e.g., hired workers reveal their productivity by the number of units produced on the job, which is unobserved among unhired workers. The interaction component—and hence systemic and total discrimination—will typically be only partially identified in this case, though additional quasi-experimental

variation in the outcome selection mechanism can bridge this gap (see, e.g., Arnold et al. 2022; Baron et al. 2024; Rambachan et al. 2024). In other settings, Y_i^0 may not even be selectively observed and must be proxied by other observables. Here, frameworks like those in Altonji et al. (2005) or Oster (2017) may be used to gauge the sufficiency of the proxies for capturing systemic and total discrimination. As before, we expect the experimental IA approach to be more tractable in these extensions by reducing the number of conditional-on-Y_i^0 objects to be estimated.

5 Constructive IA Application

Our first field experiment uses the constructive IA approach to explore the impact of racial discrimination in initial work experience on later hiring outcomes.[42] We focus on a two-node subsystem in which the first node generates an action (experience) that is part of the signal at the second, focal node (a later hiring decision). At the first node, Black and White workers apply to a set of jobs with no prior work experience. They are either hired for one of the jobs or not. They then apply for a job at the focal node either with or without prior experience from the first node, depending on whether or not they were hired. Direct discrimination in hiring inexperienced workers can thus generate systemic discrimination in subsequent hiring due to racial disparities in work experience. We use a correspondence study to estimate direct discrimination for experienced and inexperienced workers, and then use additional data to construct the interaction component and estimate systemic and total discrimination.

5.1 Methods

We first measured direct discrimination in callback rates for inexperienced versus experienced applicants at a national sample of automotive firms. We targeted entry-level positions in the automotive sector because it had previously been shown to exhibit sizable racial discrimination in callback rates (Kline et al. 2022).[43] Following a standard correspondence study design, we generated sets of fictitious resumes that varied in the applicant's race and non-race attributes (e.g., their address, high school). Race was randomly varied through the applicant's name, which was either distinctively White or Black following the literature (e.g., Bertrand and Mullainathan 2004). Job experience was also randomly varied by whether the resume included relevant experience, corresponding to one prior job in the automotive sector, or not. The resumes were submitted to online job postings and subsequent callbacks were recorded. See Appendix C.1.1-Appendix C.1.4 for details on the experimental design.

[42] See https://aspredicted.org/5XT_YPY for pre-registration materials. Per the pre-registration, we first ran a pilot to determine the feasibility of the experimental design. There were no qualitative differences in any of the comparisons between the pilot and full study. As outlined in the document, we pool the data for the main analysis.

[43] We used the same sample of automotive firms as in Kline et al. (2022). Notably, our experiment was completed prior to the public identification of those firms in Kline et al. (2024).

To estimate the likelihood that a callback translated to a job offer, and hence experience, we obtained a measure of the typical conversion rate for callbacks. Specifically, we used a recruiting agency to survey 107 hiring managers with experience in evaluating applicants for entry-level jobs in the automotive industry. We elicited the managers' estimated probability that a callback would lead to a job offer as a function of the applicant's previous experience and race. See Appendix C.1.5 for details on the survey.

Mapping this data to our framework, consider the two-node subsystem $\mathcal{N}^0 = \{1, n^*\}$ where node 1 corresponds to an initial job search and focal node n^* corresponds to a hiring decision in a subsequent job search. At node 1, worker i applies to J_i jobs with resume S_i^1 (which does not contain any relevant experience) and receives hiring decision $A_i^1 = (A_i^{1,1}, ..., A_i^{1,J_i})$, where $A_i^{1,j} = A^{1,j}(G_i, S_i^1)$ denotes whether or not the worker is offered job j. We assume that a worker can accept at most one job offer and if the worker receives at least one offer, then she accepts it and it translates to a line of experience. Therefore, this first-node hiring action A_i^1 affects the worker's second-node resume S_i^* (the node-n^* signal) by either adding a line of relevant experience or not to the first-node resume S_i^1, depending on whether the worker obtained at least one job offer or not. At node n^*, worker i applies to a job with resume S_i^* and receives hiring decision $A_i^* = A^*(G_i, S_i^*)$, which corresponds to whether or not she is offered the job. The dependence of A_i^* on A_i^1 via S_i^* captures the link between first-node hiring and second-node hiring: direct discrimination at the first node ($A^1(b, S_i^1) \neq A^1(w, S_i^1)$) can translate to different rates of experience, and hence, systemic discrimination in subsequent hiring. We set the reference qualification as the first-node resume, $Y_i^0 = S_i^1$, to isolate the systemic impact of direct discrimination in hiring inexperienced workers.

The data described above generates an estimate of the treatment component in this subsystem, i.e., direct discrimination at each node and signal. The data on hiring conversion rates allows us to translate callback rates to hiring rates at each node—for node 1, the average hiring rate $A_i^{1,j}$ across workers with no experience, and for node n^*, the average hiring rate A_i^* across workers with each level of experience. Specifically, multiplying the average callback rate and average callback conversion rate for a given race and experience level yields an estimate of the corresponding job offer rate. Direct discrimination in first-node hiring then corresponds to the difference between the job offer rate for White versus Black workers with no relevant experience. This is also the estimate of direct discrimination in second-node hiring for workers who failed to gain experience at the first node. Direct discrimination in second-node hiring for experienced workers is the analogous difference for White versus Black workers with relevant experience from the first node.

To estimate the interaction component, i.e., how direct discrimination in first-node hiring indirectly affects second-node hiring through listed job experience, we also need a measure of the likelihood of gaining experience at the first node. This depends on the number of

jobs the applicant applies to. We proxy this number with data on the number of local job openings (i.e., local labor market thickness). Specifically, we scraped the average number of entry-level job openings at automotive firms over a week for each municipality included in the experiment. See Appendix C.1.6 for details. This data, together with the job offer rate estimates described above, yields an estimate of the probability of becoming experienced at the second node by race.[44]

We combine these components to construct estimates of systemic and total discrimination, following Section 4.2. We average across qualification levels $Y_i^0 = S_i^1$ (i.e., first-node resumes), as the design imposes the same qualification distribution across race.

5.2 Results

We first analyze the callback rates in the experiment, as reported in Table 1. Column 1 shows sizable direct discrimination in callback rates against Black applicants. Overall, 28.3% of White applicants received a callback compared to 20.0% of Black applicants—a significant 8.3 percentage point gap ($p < 0.01$). Columns 2 and 3 further show substantial heterogeneity with respect to prior job experience. Among resumes without experience, callback rates were 25.6% for White applicants and 13.2% for Black applicants, resulting in a significant 12.4 percentage point disparity ($p < 0.01$). This gap is larger than callback disparities found in previous studies, which, to the best of our knowledge, included previous work experience on the resumes (Bertrand and Mullainathan 2004; Nunley et al. 2015; Deming et al. 2016). Indeed, experience shrank the racial disparity to 4.2 percentage points, with 31.1% of White applicants and 26.9% of Black applicants receiving a callback. While not statistically significant ($p = 0.20$), this disparity roughly matches the 5.3 percentage point gap ($p < 0.01$) that Kline et al. (2024)—who only used applications with experience—find for the same types of automotive sector jobs. Column 4 summarizes the return to experience: among all applicants, experienced resumes received callbacks at a 9.5 percentage point (49%) higher rate than inexperienced resumes ($p < 0.01$).

Turning to callback conversion rates, hiring managers reported that, on average, 55% of applicants without experience received a job offer, compared to 71% of applicants with experience. This difference is statistically significant ($p < 0.01$). The racial disparity in reported callback conversion rates—approximately 3 percentage points—was not significant. These relatively high conversion rates suggest that the racial disparities in callback rates will translate into meaningful disparities in job offers.

Finally, local market thickness was moderate and varied across municipalities: the 25^{th},

[44]Specifically, we assume the probability of being offered each job is independently distributed for inexperienced workers conditional on race. Letting p_g denote this probability for race g, the probability of receiving no job offers is $(1 - p_g)^J$, where J corresponds to the number of local job openings, and the probability of gaining experience is $1 - (1 - p_g)^J$.

TABLE 1. Constructive IA Application: Effects of Race and Experience on Callback Rates

	(1) Full Sample	(2) Inexperienced	(3) Experienced	(4) Full Sample
Black	-0.083***	-0.124***	-0.042	
	(0.022)	(0.030)	(0.033)	
Experienced				0.095***
				(0.021)
Constant	0.283***	0.256***	0.311***	0.194***
	(0.023)	(0.028)	(0.030)	(0.020)
Observations	1,001	500	501	1,001

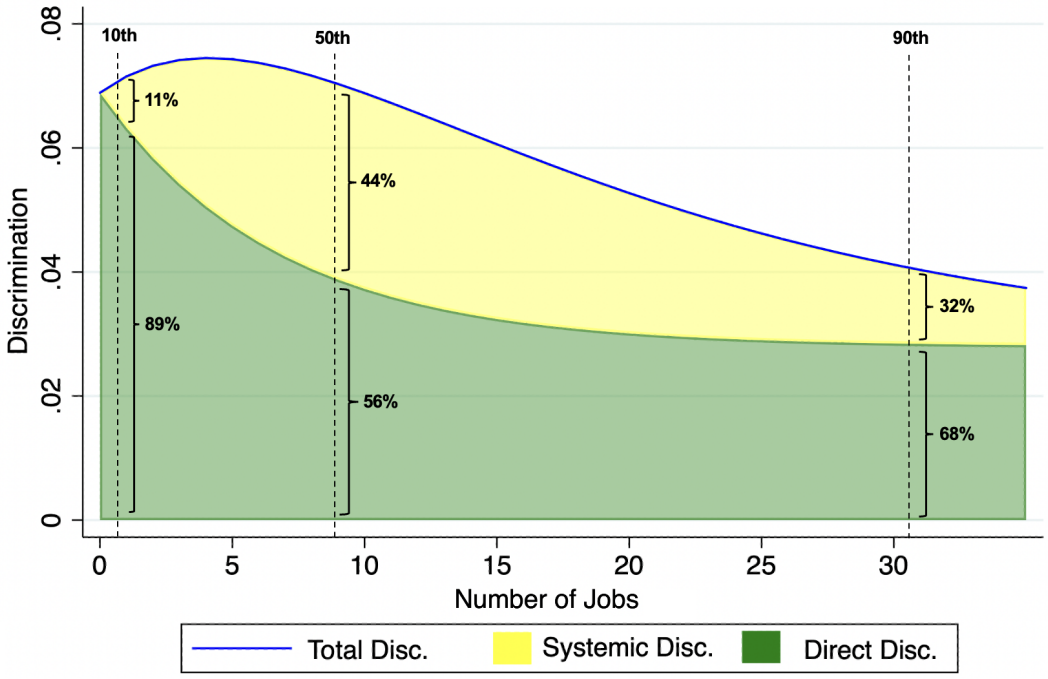
Notes: This table reports estimated coefficients from regressing a callback indicator on either an indicator for a resume having a distinctively Black name or a resume having prior relevant job experience. Standard errors, clustered at the posting level, are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

50^{th}, and 75^{th} percentiles of number of job postings were 5, 8, and 18, respectively. The moderate number of postings suggests that the relatively large racial disparity in job offer rates for inexperienced workers will lead to notable disparities in second-node experience, since applicants do not have a large number of chances to receive an offer. Moreover, the variability in local market thickness suggests that the systemic discrimination stemming from this experience disparity will vary meaningfully across regions.

Figure 5 shows estimates of total, systemic, and direct discrimination at the second (focal) node as a function of the number of first-node job openings. In markets with a median thickness of 8 openings, White applicants are around 7.5 percentage points more likely to be hired than Black applicants, after having had a chance to secure prior job experience. More than half of this total discrimination is due to systemic discrimination against Black applicants, who are less likely to have valuable prior experience due to direct discrimination in first-node hiring. The remainder is due to direct discrimination in second-node hiring, where White applicants are more likely to be hired than Black applicants with the same experience. A standard correspondence study—which conditions on experience—would therefore miss nearly half of the total discrimination in this subsystem.

These estimates demonstrate that market conditions (here, local labor market thickness) can play a key role in amplifying or mitigating the impact of past discrimination. When the market is very thin or thick, direct discrimination in first-node hiring will not generate large differences in experience. Intuitively, if the market is thin with very few openings, then few applicants gain experience, regardless of race. Most enter the second node without experience, and direct discrimination towards inexperienced Black workers drives total discrimination. At the other extreme, in thick markets with many openings, applicants

FIGURE 5. Constructive IA Application: Discrimination Decomposition



Note: This figure shows total discrimination $\Delta^T(y^0)$, average direct discrimination $\Delta^D(y^0, w)$, and systemic discrimination $\Delta^S(y^0, b)$ as a function of local market thickness, averaged across qualification levels y^0.

have many chances to be hired and gain experience. Therefore, the hiring decision for any one opening has little impact on future experience—most applicants land a job and enter the second node with experience. Hence direct discrimination towards experienced Black workers drives total discrimination. In contrast, in markets of intermediate thickness, direct discrimination in first-node hiring can translate to large differences in experience. In our data, systemic discrimination is 11% of total discrimination in thin markets with only one job posting (10^{th} percentile) and 32% in thick markets with 31 jobs (90^{th} percentile), but increases to 44% in intermediate markets with 8 jobs (50^{th} percentile).

From both a conceptual and policy perspective, this analysis suggests that market thickness is an important factor when interpreting disparities in callback or hiring rates. First, it provides a broader lens for interpreting prior results on callback and interview rate disparities by showing how they translate into labor market outcomes. Second, the analysis informs the efficacy of potential interventions aimed at reducing systemic discrimination: targeting initial disparities is most important in markets of intermediate thickness.

It is worth noting that, while we illustrate the approach in a simple two-node system, our findings suggest that systemic discrimination may be even more pronounced when work experience is affected by direct discrimination across more than two hiring nodes. In this case, conventional correspondence studies could miss an even larger share of total discrimination

by conditioning on all prior work experience.

6 Experimental IA Application

Recent studies find significant disparities in how male and female candidates are described to potential employers—the recommendation letters written for similarly qualified male versus female candidates differ significantly in the language used (Trix and Psenka 2003; Schmader et al. 2007). But whether and how such differences in language contribute to disparities in labor market outcomes is not obvious. For example, hiring managers may focus on 'hard' information and ignore language differences; alternatively, the differences may lie in language that is not relevant for the hiring decision.

We use the experimental IA approach to explore the potential impact of language on hiring outcomes in a labor market for STEM positions. Building on the example discussed in Section 4.3, we focus on a two-node hiring subsystem in which the first node generates an action (the recommendation letter) that is part of the signal at the second focal node (the hiring decision).[45] At the first node, men and women with similar resumes are evaluated and receive a recommendation letter. They then apply for a job at the second (focal) node with these letters and their resumes. Direct discrimination in the content and style of the recommendation letter (i.e., gendered language and tone) can thus generate systemic discrimination in hiring. Applying the experimental IA approach, we simulate letters for male versus female applicants and elicit hiring outcomes and wage offers for the three groups outlined in Figure 4. This yields estimates of systemic and total discrimination in hiring due to gender disparities in recommendation letters. This experimental design avoids the challenges of estimating the interaction component in the constructive IA approach, which would be prohibitive given the high dimensionality of the recommendation letter signal.

6.1 Methods

For the first node, we used a large language model (LLM) to generate a set of recommendation letters for fictitious job applicants with recent college degrees in science, technology, engineering, or mathematics (STEM). LLMs have been shown to generate gendered language that matches differences in letters written by humans (Wan et al. 2023). We generated a set of resumes that varied in the applicant's gender, as signaled through name (distinctively male or female). These resumes were otherwise identical in terms of relevant education and experience. We then prompted the LLM to generate a recommendation letter for each resume. This approach lets us hold fixed the relevant qualifications of candidates while replicating the gendered language and tone employed by recommendation letter writers. Indeed, the

[45] See https://aspredicted.org/w4g6-kzgq.pdf for pre-registration materials. We pre-registered three sets of materials based on the design in Figure 4. As outlined below, we ran a fourth group to check robustness to the alternative decomposition.

letters in our study exhibited similar gender differences—on dimensions such as individual agency, leadership ability, and communality—as previously documented for both LLMs and humans (Trix and Psenka 2003; Schmader et al. 2007).[46] See Appendix C.2 for details.

For the second (focal) node, we used a professional survey targeting agency to recruit 396 hiring managers. These managers had experience in evaluating applicants for STEM jobs that typically require a recommendation letter. Each manager observed the resumes and recommendation letters for a random set of two applicants. They then reported the likelihood that they would recommend each applicant to the next stage of the recruitment process (on a scale of 1 to 10) and each applicant's expected hourly wage should they be hired. These reports were incentivized using a similar methodology to Kessler et al. (2019): the managers were aware that the applicants were fictitious, but the attributes (e.g., prior work experience) could be matched to actual applicants with similar attributes. Managers were shown matched real applicants based on the fictitious applicants that they rated highly.[47]

Applicants were drawn from four groups, mirroring the groups illustrated in Figure 4. The endogenous-m and exogenous-f groups were similar to those which would be used in a standard correspondence or audit study: the resumes and recommendation letters were identical except for using distinctively male or female names and pronouns. Specifically, the resumes differed only in the candidate's name, and the recommendation letters were generated from this resume with a male name; for the exogenous-f group, the name and pronouns were then changed to be distinctively female. The endogenous-f group had resumes with distinctively female names, but differed from the exogenous-f and endogenous-m groups in that the recommendation letters were generated from the resume with a female name. The fourth group—exogenous-m—had resumes with distinctively male names and recommendation letters generated from the resume with a female name. Adding this group enabled us to also measure systemic discrimination under the male action rule.

Mapping this data to our framework, given a two-node subsystem $\mathcal{N}^0 = \{1, n^*\}$, node 1 corresponds to obtaining a recommendation letter and focal-node n^* corresponds to a hiring decision. The first-node action $A_i^1 = A^1(G_i, S_i^1)$ corresponds to the letter for worker i of gender G_i based on resume S_i^1. This action is part of the focal-node signal $S_i^* = (A_i^1, S_i^1)$, which consists of the recommendation letter and resume. Direct discrimination at the first node ($A^1(f, S_i^1) \neq A^1(m, S_i^1)$, where we set $G_i \in \{m, f\}$ to denote male and female) corresponds to using systematically different language when writing letters for male versus female workers with the same resume. This direct discrimination can generate systemic

[46] For example, "Matthew independently spearheaded projects to create..." versus "Emily reliably developed and maintained software applications...", or "Jacob Meyer is a self-driven, highly capable professional..." versus "Mary is a diligent professional with strong communication skills..."

[47] This factorial design is known as an Incentivized Resume Rating paradigm. See Lahey and Oxley (2021) and Kübler et al. (2018) for similar uses of factorial designs in studying discrimination.

discrimination in focal-node hiring and wage actions $A_i^* = A^*(G_i, S_i^*)$, given the dependence of S_i^* on A_i^1. We set the reference qualification as the resume, $Y_i^0 = S_i^1$, to isolate the systemic impact of direct discrimination in letter writing.

Critically, relative to the study in Section 5, the component of signal S_i^* generated in the subsystem (i.e., the letter) is high-dimensional. This precludes measuring the impact of each signal realization, which would be required for estimating the interaction component in the constructive IA approach. Instead, the experimental approach constructs measures of systemic and total discrimination from a comparison of the four groups outlined above. Specifically, comparing the hiring outcomes of the endogenous-m vs. exogenous-f groups identifies average direct discrimination under the male signal distribution, comparing exogenous-f vs. endogenous-f identifies systemic discrimination under the female action rule, and comparing endogenous-m vs. endogenous-f identifies total discrimination. Analogously, the exogenous-m group together with endogenous-m and endogenous-f allows us to measure systemic discrimination under the male action rule. The analysis that follows focuses on the first case, with the second case presented below for robustness.

6.2 Results

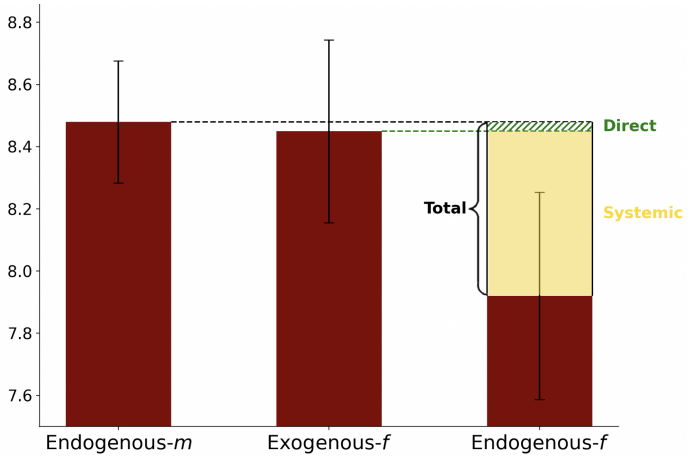
We found substantial total discrimination against female applicants. Figure 6 shows that the endogenous-f group had significantly lower average hiring likelihood and prospective wage than the endogenous-m group, corresponding to roughly 29% and 31% of the respective outcome standard deviations ($p < 0.01$ for both outcomes). In contrast, we found small and insignificant levels of direct discrimination. The average hiring likelihood for the exogenous-f group was only slightly below that of the endogenous-m group, while exogenous-f had a slightly *higher* average prospective wage (though this gap is not statistically significant). As discussed below, these results are consistent with recent work finding minimal direct gender discrimination in audit and correspondence studies.

As shown in Figure 6, this significant total discrimination is driven by systemic discrimination that stems from gendered differences in the language used in the recommendation letters. The endogenous-f group—who had letters written for female applicants—had a substantially lower hiring likelihood and prospective wage than the exogenous-f group—who had letters written for male applicants. These gaps represent 27% and 43% of the respective outcome's standard deviation and the vast majority (or more) of total discrimination.[48]

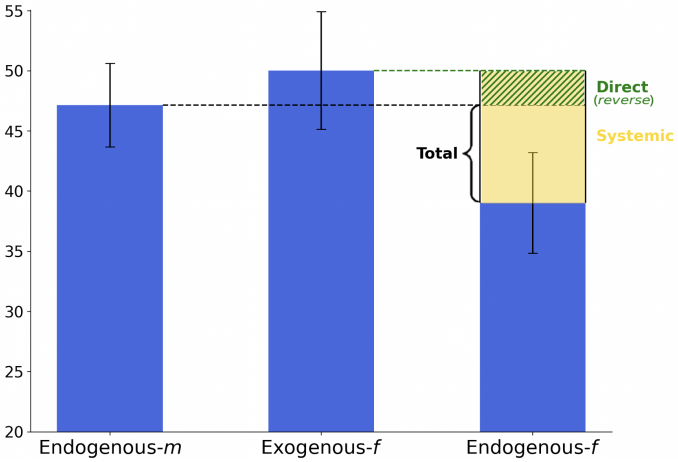
To demonstrate the utility of the experimental IA approach in this setting, we consider an alternative analysis that uses low-dimensional statistics to summarize the language of the high-dimensional recommendation letter signal. Specifically, we follow Kaplan et al. (2024) in

[48]The alternative decomposition yielded similar findings. Both the hiring likelihood and prospective wage for the exogenous-m group (7.36 and 38.85, respectively) was similar to the endogenous-f group, indicating little direct discrimination. Thus, this decomposition also shows that most total discrimination is systemic.

FIGURE 6. Experimental IA Application: Results



(a) Hiring Likelihood



(b) Prospective Wage

Note: This figure shows average hiring likelihoods and prospective wages for the three experimental IA groups, as well as the resulting estimates of total, direct, and systemic discrimination. The error bars report 95% confidence intervals.

summarizing letter language across the four dimensions that have been found to be strongly gendered in recommendation letters (Schmader et al. 2007). These indices measure language associated with ability, agency, leadership, and being a standout job candidate.

If these indices successfully captured the key aspects of language responsible for generating the observed hiring and wage disparities, then one could potentially use the constructive IA approach with text data via such summary statistics. Table 2 shows that this would be challenging here: controlling for this set of indices has little impact on the measured gender gap in hiring likelihoods or prospective wages between the endogenous-m and endogenous-f groups. In the case of the hiring likelihood outcome, the controls lead to a small decrease in the gender gap; in the case of prospective wages, the gender gap actually increases. In both

TABLE 2. Experimental IA Application: Effects of Gender and Language Constructs

	Hiring Likelihood		Prospective Wage	
	(1)	(2)	(3)	(4)
Female	-0.55***	-0.50**	-8.11***	-12.80***
	(0.16)	(0.23)	(1.73)	(3.11)
Ability		-33.30		853.27
		(44.53)		(787.32)
Leadership		22.27		19.02
		(30.46)		(428.51)
Standout		4.47		-1266.07
		(54.30)		(1083.74)
Agentic		0.02		59.88
		(24.33)		(460.43)
Observations	441	441	441	441

Notes: This table reports estimated coefficients from regressing either hiring likelihood or prospective wage on a female indicator and four indices summarizing recommendation letter language. Standard errors, clustered at the hiring manager level, are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

cases, none of the four indices are statistically significant. In other words, the low-dimensional summaries of language account for a minimal share of what drives systemic discrimination in this setting.[49] This highlights the difficulty of examining the impact of complex signals such as text data using the constructive IA approach. In our setting, without the experimental IA design, a researcher could erroneously conclude that gendered language disparities do not contribute to hiring discrimination.

Substantively, these findings speak to recent work which argues that direct discrimination is unlikely to fully explain gender disparities in the labor market. For example, a meta-analysis of correspondence studies in Schaerer et al. (2023) found little discrimination overall when non-group attributes (i.e., the signal) for male and female candidates are held fixed (although there was substantial heterogeneity across employers, with some favoring women and others favoring men; see also Kline et al. 2022). Our findings suggest that, by controlling for non-group attributes, such estimates may miss a substantial degree of gender discrimination from "soft" signals in verbal recommendations. These complex soft signals may ultimately be responsible for driving a substantial share of observed gender disparities in labor markets. This provides further evidence for the importance of relying on "hard" signals in the evaluation process.[50]

[49]For example, systemic discrimination could arise from complex interactions between different components of the letter or difficult-to-quantify aspects of the text.

[50]See Amer et al. (2024) for evidence that personal interaction—which has a similarly complex signal space as recommendation letters—may be responsible for gender bias in the technology industry.

7 Conclusion

Several large literatures, both within and outside of economics, emphasize the importance of systemic factors in driving group-based disparities. But empirical analyses largely focus on direct discrimination as a function of group identity itself. We bridge this gap by developing new theoretical and empirical tools to study and measure systemic discrimination. Our general framework formalizes a notion of systemic discrimination as arising from discrimination at other points in a system, and defines a measure of total discrimination that accounts for both direct and systemic components. We map our framework to past work on discrimination, showing that our definition of systemic discrimination unearths forces in these existing models that may have been overlooked. We then develop new empirical tools—the constructive and experimental IA designs—to identify these components in the data. Our applications demonstrate the presence of systemic discrimination in two important labor market settings, and show how conventional methods of studying direct discrimination can lead to underestimates of the total impact of discrimination in a given setting.

By formalizing the differences and possible interactions between direct and systemic discrimination, our framework can be useful for interpreting and predicting the effects of policies aimed at reducing disparities. Consider the case of racial or gender quotas. In standard models of taste-based or statistical discrimination, such policies would have a temporary effect on disparities: evaluators' decisions would revert back and the disparity would re-emerge when the quota is lifted. However, if the initial disparity was due to technological systemic discrimination, e.g., in access to skill development, then quotas may reduce the disparity in the skill distribution. De Sousa and Niederle (2022) show that the introduction of a team quota for the minimum number of female chess players improved the performance of female chess players across the country (but not outside the country), presumably, as the authors note, because this created an incentive to invest in the skill of female chess players.

New analytic tools may also broaden the set of appropriate policy responses to observed disparities. Systemic discrimination can lead to illegal disparate impact in some settings, as was found in the landmark *Griggs v. Duke Power Co. (1971)* case. The development of robust econometric methods for measuring systemic and total discrimination can be a powerful complement to existing regulatory tools in such settings.[51] Economic models of systemic discrimination can aid the interpretation of these methods, by enriching policymakers' understanding of interactions over time and across different domains.

[51] For example, the U.S. Equal Employment Opportunity Commission (EEOC) launched nearly 600 investigations into systemic discrimination in 2020. Many employment practices EEOC flags for possible systemic discrimination are indirect (such as word-of-mouth recruitment practices), and would thus not be picked up by a conventional correspondence or audit study (see https://www.eeoc.gov/systemic-enforcement-eeoc).

References

Aaronson, D., D. Hartley, and B. Mazumder (2021): "The Effects of the 1930s HOLC 'Redlining' Maps," *American Economic Journal: Economic Policy*, 13, 355–92.

Abramitzky, R., J. Conway, R. Mill, and L. C. Stein (2023): "The Gendered Impacts of Perceived Skin Tone: Evidence from African American Siblings in 1870–1940," Working Paper 31016, National Bureau of Economic Research.

Agan, A. and S. Starr (2017): "The Effect of Criminal Records on Access to Employment," *American Economic Review P&P*, 107, 560–64.

Agan, A. Y., B. Cowgill, and L. K. Gee (2024): "Salary History and Employer Demand: Evidence from a Two-Sided Audit," *American Economic Journal: Applied Economics*, forthcoming.

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

Allard, S. W. and M. L. Small (2013): "Reconsidering the Urban Disadvantaged: The Role of Systems, Institutions, and Organizations," *Annals of the American Academy of Political and Social Science*, 647, 6–20.

Althoff, L. and H. Reichardt (2024): "Jim Crow and Black Economic Progress after Slavery," *The Quarterly Journal of Economics*, 139, 2279–2330.

Altonji, J. G. and R. M. Blank (1999): "Race and gender in the labor market," *Handbook of Labor Economics*, 3, 3143–3259.

Altonji, J. G., T. E. Elder, and C. R. Taber (2005): "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools," *Journal of Political Economy*, 113, 151–184.

Amer, A., A. C. Craig, and C. Van Effenterre (2024): "Decoding Gender Bias: The Role of Personal Interaction," IZA Discussion Paper 17077, Institute of Labor Economics.

Angwin, J., J. Larson, S. Mattu, and L. Kirchner (2016): "Machine Bias," *ProPublica Report*.

Arnold, D., W. Dobbie, and P. Hull (2021): "Measuring Racial Discrimination in Algorithms," *AEA Papers and Proceedings*, 111, 49–54.

——— (2022): "Measuring Racial Discrimination in Bail Decisions," *American Economic Review*, 112, 2992–3038.

Arnold, D., W. Dobbie, and C. S. Yang (2018): "Racial Bias in Bail Decisions," *Quarterly Journal of Economics*, 133, 1885–1932.

Arrow, K. J. (1973): "The Theory of Discrimination," in *Discrimination in Labor Markets*, ed. by O. Ashenfelter and A. Rees, Princeton, NJ: Princeton University Press.

Avery, M., A. Leibbrandt, and J. Vecci (2023): "Does Artificial Intelligence Help or Hurt Gender Diversity? Evidence from Two Field Experiments on Recruitment in

Tech," Monash Economics Working Papers 2023-09, Monash University, Department of Economics.

Baron, E. J., J. J. Doyle Jr, N. Emanuel, P. Hull, and J. P. Ryan (2024): "Discrimination in Multiphase Systems: Evidence from Child Protection," *The Quarterly Journal of Economics*, 139, 1611–1664.

Barron, K., R. Ditlmann, S. Gehrig, and S. Schweighofer-Kodritsch (2024): "Explicit and Implicit Belief-Based Gender Discrimination: a Hiring Experiment," *Management Science*, forthcoming.

Bartik, A. W. and S. T. Nelson (2024): "Deleting a Signal: Evidence from Pre-Employment Credit Checks," *The Review of Economics and Statistics*, 1–47.

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

Berk, R., H. Heidari, S. Jabbari, M. Kearns, and A. Roth (2021): "Fairness in Criminal Justice Risk Assessments: The State of the Art," *Sociological Methods & Research*, 50, 1–42.

Bertrand, M. and E. Duflo (2017): "Field Experiments on Discrimination," in *Handbook of Economic Field Experiments*, ed. by A. V. Banerjee and E. Duflo, North-Holland, vol. 1, 309–393.

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

Bierer, B. E., L. G. Meloney, H. R. Ahmed, and S. A. White (2022): "Advancing the inclusion of underrepresented women in clinical research," *Cell Reports Medicine*, 3.

Blank, R. M. (2005): "Tracing the economic impact of cumulative discrimination," *American Economic Review*, 95, 99–103.

Blinder, A. S. (1973): "Wage Discrimination: Reduced Form and Structural Estimates," *Journal of Human Resources*, 8, 436–455.

Bohren, J. A., K. Haggag, A. Imas, and D. G. Pope (2023): "Inaccurate Statistical Discrimination: An Identification Problem," *Review of Economics and Statistics*, 1–45.

Bohren, J. A., A. Imas, and M. Rosenberg (2019): "The Dynamics of Discrimination: Theory and Evidence," *American Economic Review*, 109, 3395–3436.

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

——— (2019): "Beliefs About Gender," *American Economic Review*, 109, 739–73.

Brekoulakis, S. (2013): "Systemic Bias and the Institution of International Arbitration: A New Approach to Arbitral Decision-Making," *Journal of International Dispute Settlement*, 4, 553–585.

Buchmann, N., C. Meyer, and C. D. Sullivan (2023): "Paternalistic Discrimination,"

Working Paper.

Buolamwini, J. (2022): "Facing the Coded Gaze with Evocative Audits and Algorithmic Audits," Ph.D. thesis, Massachusetts Institute of Technology.

Bursztyn, L., T. Fujiwara, and A. Pallais (2017): "'Acting Wife': Marriage Market Incentives and Labor Market Incentives," *American Economic Review*, 107, 3288–3319.

Cain, G. G. (1986): "The Economic Analysis of Labor Market Discrimination: A Survey," in *Handbook of Labor Economics*, Elsevier, vol. 1, 693–785.

Ceci, S. J. and W. M. Williams (2011): "Understanding current causes of women's underrepresentation in science," *Proceedings of the National Academy of Sciences*, 108, 3157–3162.

Coate, S. and G. C. Loury (1993): "Will Affirmative-Action Polices Eliminate Negative Stereotypes?" *American Economic Review*, 83, 1220–1240.

Coffman, K. B., C. L. Exley, and M. Niederle (2021): "The Role of Beliefs in Driving Gender Discrimination," *Management Science*, 67, 3551–3569.

Cook, C., R. Diamond, J. V. Hall, J. A. List, and P. Oyer (2021): "The Gender Earnings Gap in the Gig Economy: Evidence from over a Million Rideshare Drivers," *The Review of Economic Studies*, 88, 2210–2238.

Cook, L. (2014): "Violence and Economic Growth: Evidence from African American Patents, 1870-1940," *Journal of Economic Growth*, 19, 221–257.

Cornell, B. and I. Welch (1996): "Culture, Information, and Screening Discrimination," *Journal of Political Economy*, 104, 542–571.

Darity, W. (2005): "Stratification Economics: the Role of Intergroup Inequality," *Journal of Economics and Finance*, 29, 144–153.

Darity, W. A. and P. L. Mason (1998): "Evidence on Discrimination in Employment: Codes of Color, Codes of Gender," *Journal of Economic Perspectives*, 12, 63–90.

Davison, H. K. and M. J. Burke (2000): "Sex Discrimination in Simulated Employment Contexts: A Meta-analytic Investigation," *Journal of Vocational Behavior*, 56, 225–248.

De Plevitz, L. (2007): "Systemic Racism: The Hidden Barrier to Educational Success for Indigenous School Students," *Australian Journal of Education*, 51, 54–71.

De Quidt, J., J. Haushofer, and C. Roth (2018): "Measuring and Bounding Experimenter Demand," *American Economic Review*, 108, 3266–3302.

De Sousa, J. and M. Niederle (2022): "Trickle-down effects of affirmative action: A case study in France," Working Paper 30367, National Bureau of Economic Research.

Deming, D. J., N. Yuchtman, A. Abulafi, C. Goldin, and L. F. Katz (2016): "The Value of Postsecondary Credentials in the Labor Market: An Experimental Study," *American Economic Review*, 106, 778–806.

Derenoncourt, E., C. H. Kim, M. Kuhn, and M. Schularick (2024): "Wealth of

two nations: The US racial wealth gap, 1860–2020," *The Quarterly Journal of Economics*, 139, 693–750.

Eli, S. J., T. D. Logan, and B. Miloucheva (2023): "The Enduring Effects of Racial Discrimination on Income and Health," *Journal of Economic Literature*, 61, 924–940.

Fang, H. and A. Moro (2011): "Theories of Statistical Discrimination and Affirmative Action: A Survey," in *Handbook of Social Economics*, Elsevier, vol. 1, 133–200.

Feagin, J. (2013): *Systemic Racism: A Theory of Oppression*, Routledge.

Feagin, J. R. (1977): "Indirect Institutionalized Discrimination: A Typological and Policy Analysis," *American Politics Quarterly*, 5, 177–200.

Feagin, J. R. and C. B. Feagin (1978): *Discrimination American Style: Institutional Racism and Sexism*, Prentice Hall.

Fiske, S. T. (1998): "Stereotyping, Prejudice, and Discrimination," in *The Handbook of Social Psychology*, ed. by D. T. Gilbert, S. T. Fiske, and G. Lindzey, McGraw-Hill, 357–411, 4th ed.

Fryer, R. G. and S. D. Levitt (2004): "The Causes and Consequences of Distinctively Black Names," *The Quarterly Journal of Economics*, 119, 767–805.

Gaddis, S. M. (2017): "How Black Are Lakisha and Jamal? Racial Perceptions from Names Used in Correspondence Audit Studies," *Sociological Science*, 4, 469–489.

Gallen, Y. and M. Wasserman (2021): "Informed Choices: Gender Gaps in Career Advice," *CEPR Discussion Paper No. DP15728*.

Gawai, V. P. and J. D. Foltz (2023): "Discrimination in Science: Salaries of Foreign and US Born Land-Grant University Scientists," in *Proceedings of the Agricultural and Applied Economics Association (AAEA) 2022 Annual Meeting*.

Gebru, T. (2020): "Race and Gender," in *The Oxford Handbook of Ethics of AI*, Oxford University Press, 251–269.

Glover, D., A. Pallais, and W. Pariente (2017): "Discrimination as a Self-Fulfilling Prophecy: Evidence from French Grocery Stores," *The Quarterly Journal of Economics*, 132, 1219–1260.

Grau, N. and D. Vergara (2021): "An Observational Implementation of the Outcome Test with an Application to Ethnic Prejudice in Pretrial Detentions," Working Paper wp514, University of Chile, Department of Economics.

Gynter, P. (2003): "On the Doctrine of Systemic Discrimination and its Usability in the Field of Education," *International Journal on Minority and Group Rights*, 10, 45–54.

Hardt, M., E. Price, and N. Srebro (2016): "Equality of Opportunity in Supervised Learning," in *Proceedings of the 30th Conference on Neural Information Processing Systems*, 3323–3331.

Harrington, E. and H. Shaffer (2023): "Brokers of Bias in the Criminal System: Do

Prosecutors Compound or Attenuate Disparities Inherited from Arrest?" Working Paper.

Hill, R. B. (1988): "Structural Discrimination: The Unintended Consequences of Institutional Processes," in *Surveying Social Life: Papers in Honor of Herbert H. Hyman*, ed. by H. J. O'Gorman, Wesleyan University Press, 353–375.

Hübert, R. and A. T. Little (2023): "A Behavioral Theory of Discrimination in Policing," *The Economic Journal*, 133, 2828–2843.

Hull, P. (2021): "What Marginal Outcome Tests Can Tell Us About Racially Biased Decision-Making," Working Paper 28503, National Bureau of Economic Research.

Hurst, E., Y. Rubinstein, and K. Shimizu (2024): "Task-Based Discrimination," *American Economic Review*, 114, 1723–1768.

Kaplan, D. M., R. Palitsky, S. J. Arconada Alvarez, N. S. Pozzo, M. N. Greenleaf, C. A. Atkinson, and W. A. Lam (2024): "What's in a Name? Experimental Evidence of Gender Bias in Recommendation Letters Generated by ChatGPT," *Journal of Medical Internet Research*, 26, e51837.

Kasy, M. and R. Abebe (2021): "Fairness, equality, and power in algorithmic decision-making," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 576–586.

Kessler, J. B., C. Low, and C. D. Sullivan (2019): "Incentivized resume rating: Eliciting employer preferences without deception," *American Economic Review*, 109, 3713–44.

Kitagawa, E. M. (1955): "Components of a Difference Between Two Rates," *Journal of the American Statistical Association*, 50, 1168–1194.

Kline, P. M., E. K. Rose, and C. R. Walters (2022): "Systemic Discrimination Among Large U.S. Employers," *The Quarterly Journal of Economics*, 137, 1963–2036.

——— (2024): "A Discrimination Report Card," *American Economic Review*, 114, 2472–2525.

Knowles, J., N. Persico, and P. Todd (2001): "Racial Bias in Motor Vehicle Searches: Theory and Evidence," *Journal of Political Economy*, 109, 203–229.

Kohler-Hausmann, I. (2019): "Eddie Murphy and the Dangers of Counterfactual Causal Thinking about Detecting Racial Discrimination," *Northwestern University Law Review*, 113, 1163–1227.

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

Lahey, J. N. and D. R. Oxley (2021): "Discrimination at the Intersection of Age, Race, and Gender: Evidence from an Eye-Tracking Experiment," *Journal of Policy Analysis and Management*, 40, 1083–1119.

List, J. A. (2004): "The Nature and Extent of Discrimination in the Marketplace: Evidence

from the Field," *The Quarterly Journal of Economics*, 119, 49–89.

Lodermeier, A. (2023): "Racial Discrimination in Eviction Filing," Working Paper.

Lundberg, S. J. and R. Startz (1983): "Private discrimination and social intervention in competitive labor market," *The American Economic Review*, 73, 340–347.

May, A., J. Wachs, and A. Hannák (2019): "Gender differences in participation and reward on Stack Overflow," *Empirical Software Engineering*, 24, 1997–2019.

Mayhew, L. H. (1968): *Law and Equal Opportunity*, Harvard University Press.

McMillon, D. B. (2024): "What Makes Systemic Discrimination," Systemic?" Exposing the Amplifiers of Inequity," *arXiv preprint arXiv:2403.11028*.

Mengel, F., J. Sauermann, and U. Zölitz (2019): "Gender Bias in Teaching Evaluations," *Journal of the European Economic Association*, 17, 535–566.

Mocanu, T. (2023): "Designing Gender Equity: Evidence from Hiring Practices and Committees," IZA Discussion Paper 17480, Institute of Labor Economics.

National Archives and Records Administration (NARA) (2007): "Access to Archival Databases (AAD)," U.S. National Archives.

Neal, D. A. and W. R. Johnson (1996): "The Role of Premarket Factors in Black-White Wage Differences," *Journal of Political Economy*, 104, 869–895.

Nunley, J. M., A. Pugh, N. Romero, and R. A. 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, 1093–1125.

Oaxaca, R. (1973): "Male-Female Wage Differentials in Urban Labor Markets," *International Economic Review*, 14, 693–709.

Oster, E. (2017): "Unobservable Selection and Coefficient Stability: Theory and Evidence," *Journal of Business & Economic Statistics*, 37, 187–204.

Pager, D., B. Bonikowski, and B. Western (2009): "Discrimination in a Low-Wage Labor Market: A Field Experiment," *American Sociological Review*, 74, 777–799.

Pager, D. and H. Shepherd (2008): "The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets," *Annuual Review of Sociology*, 34, 181–209.

Phelps, E. S. (1972): "The Statistical Theory of Racism and Sexism," *American Economic Review*, 62, 659–661.

Pierson, E., C. Simoiu, J. Overgoor, S. Corbett-Davies, D. Jenson, A. Shoemaker, V. Ramachandran, P. Barghouty, C. Phillips, R. Shroff, et al. (2020): "A Large-Scale Analysis of Racial Disparities in Police Stops Across the United States," *Nature Human Behaviour*, 4, 736–745.

Pincus, F. L. (1996): "Discrimination Comes in Many Forms: Individual, Institutional, and Structural," *American Behavioral Scientist*, 40, 186–194.

Pinkston, J. C. (2003): "Screening discrimination and the determinants of wages," *Labour Economics*, 10, 643–658.

Powell, J. A. (2008): "Structural Racism: Building Upon the Insights of John Calmore," *North Carolina Law Review*, 86, 791.

Rambachan, A., A. Coston, and E. Kennedy (2024): "Robust Design and Evaluation of Predictive Algorithms under Unobserved Confounding," arXiv preprint arXiv:2212.09844.

Rambachan, A. and J. Roth (2020): "Bias In, Bias Out? Evaluating the Folk Wisdom," in *1st Symposium on Foundations of Responsible Computing*.

Rose, E. K. (2023): "A Constructivist Perspective on Empirical Discrimination Research," *Journal of Economic Literature*, 61, 906–923.

Roth, P. L., K. L. Purvis, and P. Bobko (2012): "A meta-analysis of gender group differences for measures of job performance in field studies," *Journal of Management*, 38, 719–739.

Rothstein, R. (2017): *The Color of Law: A Forgotten History of How Our Government Segregated America*, Liveright Publishing.

Ruebeck, H. (2024): "Causes and Consequences of Perceived Workplace Discrimination," Working Paper.

Sarsons, H. (2019): "Interpreting Signals in the Labor Market: Evidence from Medical Referrals," Working Paper.

Schaerer, M., C. Du Plessis, M. H. B. Nguyen, R. C. Van Aert, L. Tiokhin, D. Lakens, E. G. Clemente, T. Pfeiffer, A. Dreber, M. Johannesson, et al. (2023): "On the trajectory of discrimination: A meta-analysis and forecasting survey capturing 44 years of field experiments on gender and hiring decisions," *Organizational Behavior and Human Decision Processes*, 179, 104280.

Schmader, T., J. Whitehead, and V. H. Wysocki (2007): "A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants," *Sex roles: A Journal of Research*, 57, 509–514.

Sen, M. and O. Wasow (2016): "Race as a Bundle of Sticks: Designs that Estimate Effects of Seemingly Immutable Characteristics," *Annual Review of Political Science*, 19, 499–522.

Small, M. L. and D. Pager (2020): "Sociological perspectives on racial discrimination," *Journal of Economic Perspectives*, 34, 49–67.

Strack, P. and K. H. Yang (forthcoming): "Privacy Preserving Signals," *Econometrica*.

Terrell, J., A. Kofink, J. Middleton, C. Rainear, E. Murphy-Hill, C. Parnin, and J. Stallings (2017): "Gender differences and bias in open source: Pull request acceptance of women versus men," *PeerJ Computer Science*, 3, e111.

Trix, F. and C. Psenka (2003): "Exploring the color of glass: Letters of recommendation for female and male medical faculty," *Discourse & Society*, 14, 191–220.

Wan, Y., G. Pu, J. Sun, A. Garimella, K.-W. Chang, and N. Peng (2023): "'Kelly is a Warm Person, Joseph is a Role Model': Gender Biases in LLM-Generated Reference Letters," arXiv preprint arXiv:2310.09219.

Williams, J. A., T. D. Logan, and B. L. Hardy (2021): "The Persistence of Historical Racial Violence and Political Suppression: Implications for Contemporary Regional Inequality," *The ANNALS of the American Academy of Political and Social Science*, 694, 92–107.

Zafar, M. B., I. Valera, M. Gomez Rodriguez, and K. Gummadi (2017): "Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment," in *Proceedings of the 26th International Conference on World Wide Web*.

Zivin, J. S. G. and G. Singer (2022): "Disparities in Pollution Capitalization Rates: The Role of Direct and Systemic Discrimination," Working Paper 30814, National Bureau of Economic Research.

A Related Literature

Our framework builds on a large literature studying the role of systemic forces in driving group-based disparities (e.g., Pincus 1996; Pager and Shepherd 2008; Feagin 2013; Allard and Small 2013). While exact definitions vary (Small and Pager 2020), this systems-based approach distinguishes between direct discrimination, where individuals or firms treat people differently because of group identity itself, and indirect or systemic discrimination that considers the interlocking institutions or domains through which inequities propagate (Gynter 2003). In the systems-based approach, channels for observed disparities are taken as cumulative both within and across domains; discrimination is not just a product of a single individual or institution (Powell 2008). Systemic (or "structural") discrimination can be generated by the indirect relationships between outcomes and evaluations in roughly the same period, such as when discrimination in criminal justice drives unwarranted disparities in education and labor market outcomes.[52] It is also generated over time, such as when historic "redlining" practices in lending generates persistent disparities in credit access through its differential effects on generational wealth (e.g., Aaronson et al. 2021). The literature sometimes refers to the former as "side-effect" discrimination and the latter as "past-in-present" discrimination (Feagin and Feagin 1978; Gynter 2003; Feagin 2013).

Importantly, the systemic perspective shifts focus from the motives and biases of a given individual or institution to policies or institutional arrangements that contribute to *de facto* discrimination, perhaps without intent. Direct discrimination, either on the part of individuals or institutions, is inherently non-neutral: it arises from the explicit differential treatment of individuals on the basis of group identity. Systemic discrimination, in contrast, can exist in policies that are facially neutral by race, gender, or other protected characteristics (Hill 1988). For example, a lending algorithm which considers a person's zip code but does not use racial information when determining loan eligibility may be race neutral in design but discriminatory in practice. Black borrowers may be more likely to live in certain zip codes than equally creditworthy White borrowers, perhaps because of prior discriminatory policies in housing, employment, or financial markets.[53]

The distinction between direct and indirect discrimination is echoed in legal theories of disparate treatment and disparate impact (e.g., Gynter 2003; De Plevitz 2007; Brekoulakis 2013; Rothstein 2017). Under the disparate impact doctrine, a policy or practice may be deemed discriminatory if it leads to disparities without substantial legitimate justification—

[52] Powell (2008) considers systemic discrimination as driving disparities within a domain, e.g., the hiring and promotion practices within a firm or industry, and structural discrimination as driving disparities through the interaction of different systems.

[53] Note that policies that are facially neutral on protected characteristics may not be neutral in intent. Mayhew (1968) argues that some organizations may have accepted Civil Rights legislation mandating "color-blind" treatment because they were aware systemic discrimination could preserve the status quo.

as in *Griggs v. Duke Power Co. (1971)*.[54] A facially neutral practice may therefore be found to be discriminatory under this doctrine even in the absence of explicit categorization or animus. This notion of discrimination contrasts with the disparate treatment doctrine, which prohibits policies or practices motivated by a discriminatory purpose. Typically, proof of discriminatory intent is required for the finding of disparate treatment.[55]

A systemic perspective is also found in the recent literature on algorithmic unfairness (e.g., Angwin et al. 2016; Hardt et al. 2016; Zafar et al. 2017; Gebru 2020; Berk et al. 2021; Kasy and Abebe 2021; Arnold et al. 2021; Buolamwini 2022). An algorithm which does not directly use protected characteristics may nevertheless return systematically disparate outcome predictions or treatment recommendations among equally qualified individuals. The literature studies how interlocking systems of data collection, model fitting, and human-algorithm decision-making may generate such disparities.

Finally, research in the field of stratification economics proposes a systemic perspective as necessary for understanding group-based disparities because advantaged groups have an incentive to maintain them (Darity and Mason 1998; Darity 2005; De Quidt et al. 2018). Without considering the systemic interactions generating a specific outcome, as well as the incentives involved in maintaining this system, a researcher or policy maker may miss important channels through which group-based disparities persist.

Our work also adds to the long literature on direct discrimination in economics, which is typically modeled as a causal effect of group membership on treatment. Theoretical sources of direct discrimination include individual preferences or beliefs. In the canonical framework of taste-based discrimination, differential treatment emerges because individuals derive disutility from interacting with or providing services to members of a particular group (Becker 1957). In models of belief-based discrimination, differential treatment emerges because a decision-relevant statistic (such as labor market productivity) is unobserved, and there are group-based differences in beliefs about its distribution (Phelps 1972; Arrow 1973; Aigner and Cain 1977). While belief differences have traditionally been assumed to stem from true differences in the distributions, a recent literature has considered the role of inaccurate beliefs in driving direct discrimination (Bohren et al. 2023; Hübert and Little 2023; Barron et al. 2024). These differences may stem from a lack of information or biased stereotypes (Fiske 1998; Bordalo et al. 2016, 2019; Coffman et al. 2021), which again lead to causal effects of a protected characteristic on evaluations and decision-making. As noted in the paper, our work also contributes to the rich literature on non-direct discrimination in economics considering "pre-market" forces and historical injustices in driving disparities (Neal and Johnson 1996; Glover et al. 2017; List 2004; Cook 2014; Hurst et al. 2024; Sarsons 2019). Our framework

[54] See also *Dothard v. Rawlinson (1977)* and *Cocks v. Queensland (1994)*.
[55] See, e.g., *Washington v. Davis (1976)* and *McClesky v. Kemp (1987)*.

puts structure on these disparities so they can be decomposed and measured as different systemic channels. In Section 3.3 we also discuss how the model of Coate and Loury (1993) captures a specific source of systemic discrimination in our framework.

A rich empirical literature in economics has used both experimental and observational data to identify the causal effect of group identity on treatment, holding other observables constant (e.g., Bertrand and Mullainathan 2004; Fang and Moro 2011; Bertrand and Duflo 2017). In the widely-used correspondence study method, evaluators (e.g., hiring managers) are presented with information about individuals (e.g., applicants for a job), which consists of the individual's group identity and other signals of their qualifications (e.g., education level). Since everything but group identity—or a signal of this identity—is held constant in the experimental design, any differential treatment can be directly attributed to the causal effect of this variable. Recent advances in this methodology have been used to examine the dynamics of discrimination (Bohren et al. 2019) and the heterogeneity in discrimination across institutions (Kline et al. 2022).[56] A parallel empirical literature has developed tools to distinguish different economic theories of discrimination. Recent advances involve outcome tests of racial bias, in both observational (Knowles et al. 2001; Grau and Vergara 2021) and quasi-experimental data (Arnold et al. 2018; Hull 2021).

As also noted in Small and Pager (2020), the systemic perspective suggests that standard tools for measuring direct discrimination miss an important component. Efforts to model and measure causation at any particular juncture and within a specific domain can substantially understate the cumulative impact of discrimination across domains or time. We contribute to the economics literature by expanding the tools for studying such forms of discrimination. Additionally, our framework offers new interpretations for previously documented group-based disparities. For example, evidence for a reversal of direct discrimination over time—such as the ones documented in Bohren et al. (2019) and Mengel et al. (2019)—may not imply that total discrimination has been mitigated or reversed. If, as argued, biased evaluators drive initial discrimination in the pipeline, the group that ends up being favored may still face substantial total discrimination when conditioning on underlying qualifications.[57]

A growing literature in economics has examined the impact of historical direct discrimination on subsequent disparities. Cook (2014) and Williams et al. (2021) study the long-run effects of racial violence on innovation and regional inequality, respectively. Derenoncourt

[56] While Kline et al. (2022) refer to their study as estimating "systemic discrimination," this classification is not consistent with the large social science literature on systemic discrimination outlined above. Their correspondence study is designed to measure direct discrimination, formalized as the causal effects of protected characteristics in a hiring decision. We view this work as more accurately studying institutional direct discrimination.

[57] The systemic perspective also highlights the lasting impact of initial stereotypes (Bordalo et al. 2016, 2019). Even if signals become more precise and direct discrimination decreases, total discrimination can persist through systemic channels.

et al. (2024) and Eli et al. (2023) review and examine the impact of historical discriminatory practices on the evolution of the racial wealth gap.

A series of papers have built directly on our definitions and framework to measure and classify direct, systemic, and total discrimination. Althoff and Reichardt (2024) measure the systemic components of disparities that stem from racially oppressive institutions—slavery and Jim Crow laws. Baron et al. (2024) examine discrimination in foster care through the investigator-screener relationship, finding that systemic discrimination generated by screeners accounts for a substantial proportion of the resulting total discrimination. Zivin and Singer (2022) study racial differences in home values as a function of pollution exposure, concluding that 75% of the disparity was driven by systemic discrimination in complementary amenities. Lodermeier (2023) applies our framework to the study of eviction rates, finding that the substantial racial disparity is likely caused by direct rather than systemic discrimination. Gawai and Foltz (2023) look at the impact of country of birth on income in academia and find significant total discrimination. They identify two-thirds of that disparity to be driven by systemic discrimination. Finally, Buchmann et al. (2023) study a form of anticipated systemic discrimination where employers are less likely to hire women due to gender-based disparities in safety outside of the job, which they term *paternalistic discrimination*. They find that eliminating this type of discrimination would reduce the gender employment gap by 24% and increase female wages by 21% in their setting.

B Screening Discrimination

We present a theoretical example and an experiment to illustrate how group-based differences in the precision of productivity signals can lead to both direct and systemic discrimination in a screening action. The former channel is through accurate statistical discrimination: the groups face different effective thresholds for the same signal realizations because of the difference in signal precision. The latter systemic channel comes from the difference in the signal distribution, accounting for the difference in thresholds. For example, if an aptitude test is designed by a dominant group it may provide more accurate information about members of that group than for a minority group; alternatively, a medical diagnostic test may only be trialed on the majority group and is thus more predictive for this group. Such disparities in screening accuracy correspond to a type of systemic discrimination: even if individuals from different groups receive the same treatment conditional on the same test result, if the system neglects developing accurate methods to screen minority groups, these groups will face systemic discrimination.

Both the theoretical example and the experiment show that canonical statistical discrimination models may not capture the full extent of (total) discrimination stemming from differences in the signaling technology. They also show how discrimination due to differences

in the signaling technology manifests in fundamentally different ways than discrimination due to differences in the prior distribution of productivity (i.e., the other source of classic statistical discrimination). When the qualification is set to current productivity, $Y_i^0 = Y_i^*$, the former can lead to both direct and systemic forms of discrimination in the current decision, while the latter only leads to direct discrimination. Finally, we show how systemic discrimination from disparities in the informativeness of signals is likely to be heterogeneous across worker productivity levels: more productive workers tend to face more systemic discrimination than less productive workers.

B.1.1 Theoretical Example

Suppose worker productivity is distributed identically within groups, $Y_i^* \sim N(0,1)$, but the manager's signal $S_i = Y_i^* + \varepsilon_i$ has a group-specific precision: $\varepsilon_i \sim N(0, 1/\eta_g)$ when $G_i = g$, with more precise signals for group w, $\eta_w > \eta_b > 0$. The distribution of S_i for a group-g worker with productivity y is $N(y, 1/\eta_g)$ and the posterior expected productivity for a worker from group g who generates signal realization s is $s\eta_g/(1+\eta_g)$. This example sets productivity as the qualification, $Y_i^0 = Y_i^*$.

Suppose the manager hires all workers whose posterior expected productivity is at or above some threshold $t \in \mathbb{R}$: $A(g,s) = \mathbb{1}\{s\eta_g/(1+\eta_g) \geq t\}$. The manager thus hires group-g workers with signal realizations $S_i \geq t(1+\eta_g)/\eta_g$. Group-$b$ workers face a higher signal threshold, since $(1+\eta_b)/\eta_b > (1+\eta_w)/\eta_w$. Therefore, there is direct discrimination against group b stemming from the higher cutoff arising from their less precise productivity signal. Specifically, group-w workers with $S_i \in (t\frac{1+\eta_w}{\eta_w}, t\frac{1+\eta_b}{\eta_b}]$ are hired but group-b workers with signals in this range are not (hiring of workers with other signals does not depend on group).

Even without the direct discrimination in signal thresholds, however, the difference in signal precision causes equally-productive workers to be hired at different rates depending on their group. For a given $y \in \mathcal{Y}$ and $g \in \{b, w\}$, systemic discrimination is captured by

$$\begin{aligned}
&E[A(g,S_i)|Y_i^* = y, G_i = w] - E[A(g,S_i)|Y_i^* = y, G_i = b] \\
&= Pr\left(S_i \geq t(1+\eta_g)/\eta_g | Y_i^* = y, G_i = w\right) - Pr\left(S_i \geq t(1+\eta_g)/\eta_g | Y_i^* = y, G_i = b\right) \\
&= \Phi\left(\eta_b\left(t(1+\eta_g)/\eta_g - y\right)\right) - \Phi\left(\eta_w\left(t(1+\eta_g)/\eta_g - y\right)\right),
\end{aligned}$$

where $\Phi(\cdot)$ gives the standard normal distribution.

Since $\eta_b \neq \eta_w$, this expression is non-zero unless $y = t\frac{1+\eta_g}{\eta_g}$. Therefore, there is systemic discrimination almost everywhere in the productivity distribution, stemming from the differential probabilities of the signal being above a given cutoff for equally productive group-w versus group-b workers.

Systemic discrimination in this screening action is heterogeneous across worker produc-

tivity levels. With $\eta_w > \eta_b > 0$, the systemic discrimination hurts group-b workers at high levels of productivity (where $y > t\frac{1+\eta_g}{\eta_g}$) and *favors* group-$b$ workers at low levels of productivity (where $y < t\frac{1+\eta_g}{\eta_g}$) since $\Phi(\cdot)$ is strictly increasing. Intuitively, having a higher signal variance makes low-productivity group-b workers more likely to have a signal above the effective threshold by chance, while high-productivity group-b workers are more likely to generate a signal below the threshold by chance.

The average level of systemic discrimination across workers depends on which of these two productivity groups is larger. In a "cherry-picking" market with $t > 0$, such that a minority of workers are hired in each group (i.e., $Pr\left(S_i \geq t\frac{1+\eta_g}{\eta_g} \mid G_i = g\right) < 0.5$), the systemic discrimination favors group-b overall. This is because there are fewer high-productivity group-b workers hurt by the higher signal variance than low-productivity group-b workers helped by it. Conversely, in a "lemon-dropping" market ($t < 0$) with most workers being hired, the systemic discrimination hurts group-b workers overall.

This theoretical example highlights how examining screening bias with only a direct measure of discrimination may miss an important component of how differential signal precision impacts total discrimination.

Similar to the case of statistical direct discrimination (e.g., Fang and Moro 2011), differential signal precision can be heterogeneous across qualification levels. Consider, for example, a hiring decision in which the signal is equal to productivity plus mean-zero noise. A noisier signal hurts high productivity workers, as it leads to a higher chance of generating a signal below the hiring threshold, but can benefit low productivity workers by increasing the chances of the generated signal exceeding the hiring threshold. In contrast, in a medical diagnostic context, all patients benefit from a more accurate signal when it leads to more accurate diagnoses, regardless of their health status.

B.1.2 Experimental Setup

Workers: 100 participants were randomly assigned to the role of Worker. Each Worker completed two sets of tasks (A and B) and provided basic demographic information including self-reported group identity G_i (either male m or female f). Each task consisted of a test of the Worker's basic math, business, and history knowledge, with 10 randomly selected questions from these subjects. A Worker's performance on each task was defined as the number of questions she answered correctly. We restrict attention to Workers with a task-A performance in $\mathcal{S}^R = \{2, 3, 4, 5, 6\}$ in order to ensure enough data for each gender.

There were no significant gender differences in Worker performance on either task. On average, Workers completed 3.57 questions correctly on task A and 3.53 questions correctly on task B. Regressing overall performance (the sum of performance on both tasks) on a male Worker indicator yields an insignificant coefficient of -0.13 ($p = 0.84$). The gender coefficient

is similarly insignificant when we regress performance on task A (0.21; $p = 0.63$) and task B (-0.34; $p = 0.35$) separately. Performance on task B was predictive of performance on task A. Regressing the latter on the former yields a coefficient of 0.36 ($p < 0.01$). Furthermore, there were no significant gender differences in this relationship: regressing task-A performance on task-B performance, gender, and their interaction yields an insignificant interaction coefficient of 0.15 ($p = 0.58$).

Recruiters: 199 participants were randomly assigned to the role of Recruiter. Each Recruiter was shown the task-A performance of two Workers, along with the Workers' gender, and asked to select which Worker they would prefer to hire. Recruiters were then paid 1 USD for each question the hired Worker answered correctly on task B, beginning after the fifth correct answer. The Recruiter's action rule is thus $A_i^R \in \{0, 1\}$.

Hiring Managers: 501 participants were randomly assigned to the role of Hiring Manager. Each saw one Worker's profile after their evaluation by a Recruiter, along with the Worker's gender. They were shown information on the Worker's task-A performance only if the Recruiter had chosen to hire them; otherwise they saw no performance information. Therefore, $\mathcal{S}^H = \{\emptyset, 2, 3, 4, 5, 6\}$. Hiring Managers then made a binary decision of whether or not to hire the Worker. If the Worker was hired, the Hiring Manager received a bonus corresponding to their task-B performance; otherwise, the Hiring Manager received 4 dollars with certainty.

Formally, each Hiring Manager j observed a signal S_i^H corresponding to Worker i's task-A performance if the Worker was hired by the recruiter ($A_i^R = 1$). If the Worker was not hired ($A_i^R = 0$), the Hiring Manager observed no signal ($S_i^H = \emptyset$). Recruiter actions thus affected the *informativeness* of Hiring Manager signals—whether or not she saw an objective signal of productivity. This setting was designed to emulate the process by which managers can obtain more accurate performance signals depending on whether potential Workers had access to prior opportunities to "prove themselves" (e.g., internships). The Manager's action $A_i^H \in \{0, 1\}$ corresponds to her hiring the Worker.

B.1.3 Results

We measure systemic and total discrimination with respect to task-A performance, $Y_i^0 = S_i^R$, with $\mathcal{Y}^0 = \{2, 3, 4, 5, 6\}$. Since this qualification measure coincides with the Recruiter signal, any discrimination in the Recruiter stage is direct. Discrimination in the Hiring Manager stage can again be direct or systemic. We expect the differences in signal informativeness to lead to heterogeneity in systemic discrimination by qualification.

Recruiters: Recruiters directly discriminated against female Workers. The hiring rate for male Workers was 28 percentage points higher than for female Workers ($p < 0.01$), who

were hired at a rate of 36%.[58] Given the lack of gender-based performance differences, as reported in Appendix B.1.2, this disparity in hiring rates is not consistent with accurate statistical discrimination. Therefore, Recruiter direct discrimination stems from either biased preferences or beliefs.

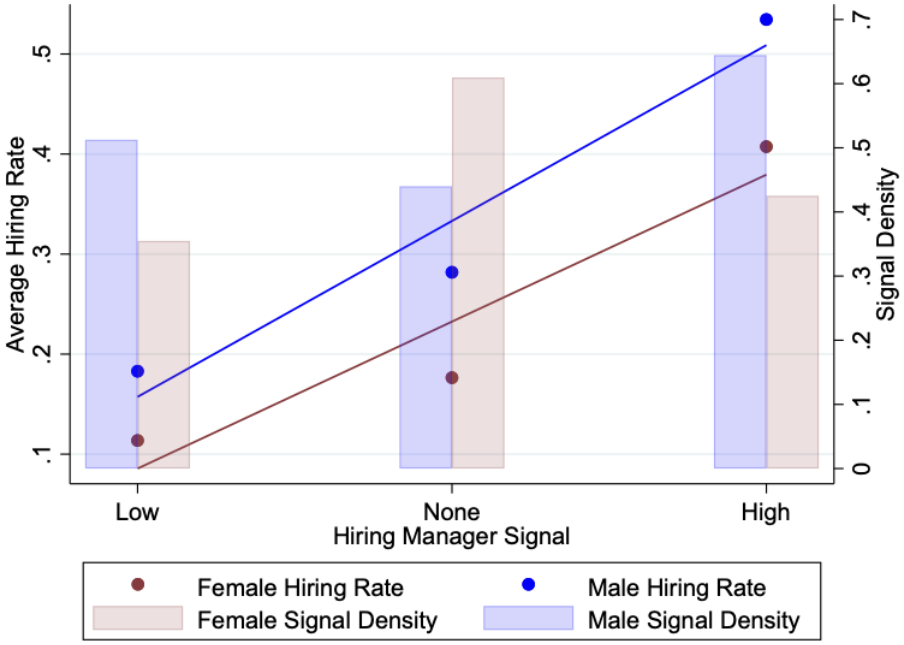
Hiring Managers: Hiring Managers discriminated against female Workers. On average, male Workers were hired at a 9 percentage point higher rate than female Workers ($p = 0.02$), who were hired at a rate of 22%. However, this average effect masks important heterogeneity. Among Workers with low (below-median) qualification levels, male Workers were hired at an insignificant 4 percentage point higher rate ($p = 0.43$).[59] Among Workers with high (above-median) qualification levels, male Workers were hired at a significant 23 percentage point higher rate ($p < 0.01$).

Figure 7 illustrates the reason for this heterogeneity in total discrimination. Similar to Figure 6, the scatter plot shows the average Hiring Manager actions conditional on the signal (or lack thereof) and the Worker's gender. The lines of best fit show a positive relationship between the signal and the probability of getting hired for both groups: Hiring Managers were more likely to hire a Worker after seeing a high signal (at or above the median score) than a low signal, with the hiring rate for no signal laying in between. Conditional hiring rates are shifted upward for male Workers, illustrating direct discrimination. Importantly, however, the distribution of signals seen by Managers also differs by gender: direct discrimination by Recruiters made Managers more likely to see both low and high signals from male Workers than from female Workers, with female Workers being much more likely to have an uninformative signal. Given the upward-sloping lines, female Workers with high qualification levels were likely to be hurt by systemic discrimination, while female Workers with low qualification levels were likely to be helped by it.

We quantify total, direct, and systemic discrimination in Hiring Manager actions using the decompositions in Section 3.2. We estimate Hiring Manager total discrimination $\Delta(y^0)$ by comparing male and female hiring rates based on task-A performance. We then estimate the Hiring Manager average direct discrimination $\overline{\tau}(w, y^0)$ faced by male Workers with a given task-A performance by averaging the gender disparities across each Hiring Manager signal realization according to the distribution each task-A performance induces over this signal. Subtracting this value from the estimate of total discrimination yields an estimate of the measure of systemic discrimination.[60] We average these measures over the distribution of task-A performance, as before, separately for Workers with low (below-median) and high (above-median) qualification levels.

[58] Standard errors are clustered at the individual level.
[59] The median task-A performance was 4.
[60] Here we use the "average" decomposition, Equation (5). The other decompositions give similar results.

FIGURE 7. Screening: Hiring Manager Hiring Rate and Signals



Notes: This figure plots average Hiring Manager hiring rates (left y-axis) and signal probabilities (right y-axis) by productivity signal type for female and male workers, where high versus low signal corresponded to either above and equal to or below the median signal (3), respectively. Gender differences in the hiring rates for a given signal illustrate direct discrimination, while gender differences in the signal probability highlight the source of systemic discrimination.

Table 3 confirms the heterogeneity in systemic discrimination faced by women with different qualification levels. For highly qualified women, total discrimination is estimated as a significant 24%. Our decomposition shows this is driven by a combination of significant direct (15%) and systemic discrimination (9%). In contrast, total discrimination among workers with a low qualification is small and insignificant (3%), despite significant direct discrimination (7%). The reason is a small degree of negative systemic discrimination among less qualified Workers (-4%). Consistent with the model in Appendix B.1.1, the gap in systemic discrimination across qualification levels is significant ($p = 0.04$).

TABLE 3. Screening: Discrimination Decomposition

	High Qualification (1)	Low Qualification (2)	Difference (3)
Total	0.24***	0.03	0.21***
	(0.06)	(0.04)	(0.07)
Average Direct	0.15***	0.07**	0.08
	(0.05)	(0.04)	(0.05)
Systemic	0.09**	-0.04	0.13**
	(0.04)	(0.03)	(0.06)
# Observations	501	501	501

Notes: This table reports estimates of each measure of discrimination in Equation (6) for Hiring Manager hiring rates, averaged by an equal-weighted distribution of task-A scores for male and female Workers in the given qualification bin, where High corresponds to above or equal to the median (3) and Low corresponds to below the median. Total discrimination is measured by the average difference in hiring rates among male versus female Workers with a given task-A score. The sample includes 501 Hiring Managers, each evaluating one Worker. Robust standard errors, obtained from a weighted bootstrap, are reported in parentheses.
*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.

C Experimental Details

C.1 Constructive IA

C.1.1 Scraping Job Listings

For the IA Study

We scraped job listings from corporate career websites of five major automotive firms. For each company, we scraped open entry-level job listings for the full-time sales position with minimal requirements (high school and no experience). In general, for each career website, we filtered on full-time status, job category, and posted time (within the past two weeks) when available. After the results were filtered, our scraping script went through each page and recorded each job listing's title and URL of its detailed job listing page.

The job requirements were checked as the script recorded the job titles and links. After each page, a list of unique job titles and their URLs (randomly selected if there was more than one listing under the same title) was created, and each URL within this list was visited to check for job requirements. The script evaluated whether the job listing qualified as entry-level by searching for keywords: first, it located all sentences/clauses containing keywords such as "Required", "High School", or "Ability"; second, it went through these sentences/clauses searching for keywords such as (variations of) "high school", "degree", and "experience." Only jobs requiring just a high school degree/GED or with no requirements would count as entry-level. We counted "requiring automotive knowledge" and "experi-

ence/degree preferred" as entry-level too, since all resumes mentioned automotive experience in the former case and the requirement was not strict in the latter case. A separate list kept all the job titles whose requirements qualified as entry-level. After each page, all job listings whose titles were in the qualified list would then be included, and job titles that were not in the list (either new jobs or disqualifying jobs) would be checked. Given the disparities in language and format these companies use in their job listings, the keywords and filters were modified on a company-by-company basis. A research assistant also monitored the scraping process and manually checked job titles. Qualified jobs were matched with their store locations either from the job listings when the script recorded them in the first place (were the addresses present), or matched by searching the store identifier through the company store locators. Eventually, we randomly selected one job listing from each store for stores with multiple job listings, such that all stores would only see one set of four resumes.

For Entry-level Jobs in General

After implementing the audit experiment, we scraped job listings of all available entry-level jobs from the same automotive companies in the cities where we applied to. The script for this part of scraping was very similar to the script we used for recording jobs for the audit experiment, except that we did not restrict job categories but instead restricted store locations. For example, we applied to stores in Houston, TX, and here we scraped all entry-level jobs (not just sales) in stores in Houston, TX, such as sales and drivers. For companies allowing us to specify a city, we searched for jobs labeled as being in a specific city; for companies that allowed searching for jobs within certain radius of a location, we chose jobs within a 5-mile radius of the city (the smallest radius). A research assistant similarly monitored the process.

C.1.2 Creating Resumes

Address

Using the store locations we recorded from the job listings, we assigned a residential address to each applicant. We used the National Address Database (NAD), restricting the sample to only contain addresses whose type is residential. For each job, we reverse geocoded the store location to latitude and longitude, and created a subsample of the residential addresses whose latitude and longitude are within $\pm$ 0.29 degrees and $\pm$ 0.37 degrees of the job location's latitude and longitude, respectively. Such degrees approximately correspond to 20 miles in distance. Further, we assigned an ordinal variable to the addresses: a value of 0 if the address is in the same city as the job location, 1 if the address is in the same state but a different city, and 2 if the address is in a different state. We first ordered this subsample by the ordinal variable, such that addresses in the same city and state were ranked the highest.

We randomly selected four distinct addresses from the top 200 addresses; if there were fewer than four addresses in this subsample, we used a larger subsample by changing the ±20 miles to ±30 miles and randomly selected four distinct addresses from the top 200 addresses. In practice, we rarely had 200 rows to randomize from, given the scarcity of residences close to the stores. Further, since this ranking was based on city and state rather than distance, our method simply prioritized addresses in the same city/state within certain distance (disciplined by latitude and longitude). After this step, there existed jobs in locations where there were not enough residences, and we excluded these jobs to avoid very small towns.

Prior Employment

Two of each set of resumes had prior related job experiences. We assigned previous employers to such resumes using the 2022 InfoGroup business data. We filtered the data by primary SIC code to include only companies in the automotive industry. We also relabeled companies such as *Carquest* who experienced mergers with companies in the scraped jobs to avoid applying with job experience from the same companies. The next steps were identical to the steps in assigning nearby residences: we first looked at companies within ± 20 miles of the job locations, assigned an ordinal variable based on city/state, and randomly selected four employers from the top 200 nearby companies ranked by the ordinal variable (that were not the same as the job listing company). If there were not enough companies within ± 20 miles, the same process was repeated for the subsample within ± 30 miles. Similarly, we rarely encountered jobs with over 200 other employers nearby. Jobs with fewer than four other employers nearby were also excluded.

The start and end dates of prior employments were randomized. For each applicant with prior experience, the length of prior experience (in years) was randomly drawn from a uniform distribution $[0.5, 2]$ and rounded to the first decimal point. Correspondingly, we counted the days between the day we made the resumes and the applicant's high school graduation date, and subtracted the length of prior experience (in days) to get the applicant's unemployment days. These unemployment days were then randomly split into before and after the prior employment. The proportion of the unemployed days before the prior employment was randomly selected from uniform $[0, 1]$. The employment start date was then set as the high school graduation date plus the number of unemployment days before prior employment, and the employment end date was set as the start date plus the length of the prior employment (in days).

High School

All of our applicants were high school graduates with automotive-related workshop experience in high school. Only the names of such workshops were mentioned as part of the

high school experiences without any description: "Automotive Technology Essentials," "Automotive Diagnostics and Service", "Car Care and Repair," and "Automotive Technology Workshop." All four applications to the same job used different names for the workshop to avoid suspicion. To assign each resume a high school, we used the National Center for Education Statistics (NCES) public school data. We further filtered this public school data to contain schools offering the 12th grade as the highest grade, whose type is either regular or technical school, and whose school names are not suggestive of online or art schools. For each job, we constructed a subsample of high schools whose zip code is within ± 100 of the job location zip code. We used ± 100 to locate high schools reasonably close to the store locations. For example, for a zip code of 60637 the range would be $[60537, 60737]$. Similar to the residential addresses, we then assigned an ordinal variable to each high school depending on whether it is in the same city and/or state. Ranking the high schools using this ordinal variable, we randomly selected four distinct high schools from the top 200 schools. Likewise, we rarely ran into job locations with over 200 schools nearby to randomly select from, and we prioritized schools in the same city/state. This step also ruled out some job listings with fewer than four high schools nearby.

The duration of high school was randomized. For each applicant, we first randomly selected an integer from uniform $[17, 19]$ as the age when this applicant graduated from high school. We added this age to the applicant's birth year, and added a randomly selected number of days from uniform $[170, 220]$ to the first day of that year as the end date to ensure the end date was in June, July, or August. Similarly, for high school start date, we subtracted 3 from the graduation year, and added to the first day of that year a randomly selected number of days from $[210, 250]$ to ensure the start date was in August or September.

Name

To create the treatment by race, we used the race-salient names and surnames from Kline et al. (2024). We created a dataset with all combinations of the first names and surnames for each race, and randomly selected distinct names from these for each company. The number of randomly selected names depended on the number of job listings each company had, such that no company would see the same name more than once. This is because all companies we applied to used a central online application system and we were unsure whether duplicated names would cause potential issues.

Email

All resumes had the applicants' emails listed. To create corresponding emails for the fictitious applicants, we purchased domains and hosted emails ourselves. All applicants' emails ended with "@mailprofessional.live" and "@voyagemail.pro". The username for each appli-

cant was randomly chosen from four potential formats: first-name-surname and a random integer, surname-first-name and a random integer, first-name-initial-surname and a random integer, and surname-first-name-initial and a random integer. All of the integers were smaller than 10,000. If a username was too long, it would be replaced with a shorter version by randomly selecting from the last two formats only, and by randomly selecting a smaller number. After these email addresses were constructed, we first registered emails manually and another scraping script later registered all these emails on the hosting platform. A research assistant manually checked the emails. In some resume formats, the emails were too long and they were broken across two lines with a hyphen in between. These email addresses, with hyphens added, were also registered accordingly such that the email addresses were valid.

Phone Number

Same as email addresses, all resumes also had a phone number listed. To record the treatment effect, we assigned a unique phone number to each name we created. The phone numbers were purchased from a phone bank firm, and were set up such that all calls and texts were redirected to a single number that was monitored. We first purchased the phone numbers manually and another scraping script was later used to make the purchases. The script randomly selected one from the states' area codes to filter the phone numbers before making the purchases, to avoid too many numbers sharing the same area code. The purchased phone numbers were also manually checked and randomly tested by a research assistant. After the phone numbers were purchased and assigned to the names, they were listed in a randomly selected format: either "XXX XXX XXXX" or "(XXX) XXX-XXXX".

Volunteer Experience

We included one volunteer experience for every fictitious applicant. The volunteer locations were "Senior Center Kitchen", "Community Food Bank", "Soup Kitchen", and "Community Meal Program", and the responsibilities only included preparing and distributing food and cleaning for all locations. For each job we applied to, each fictitious resume used one of the four volunteer locations to avoid suspicion.

Template

For each job we applied to, the four fictitious applicants used four different resume templates. These templates were written in LaTeX.

C.1.3 Filing Online Applications

Birthday and Age

Each fictitious applicant had a high school graduation age between 17 and 19. If the

applicant had no prior employment, we assumed their age last year was the same as the graduation age; if the applicant had prior employment, we assumed the applicant's age last year was the graduation year plus the length of their work experience in years. To determine the applicant's birthday, we first calculated their birth year by subtracting from the year when the experiment was run their current age (last-year age plus 1). Then, we added a randomly selected number of days from uniform $[0, 364]$ to the first day of their birth year to arrive at their exact birth date. This method ensured all applicants were over 18 at the time of application, preventing age limits from confounding the results.

SSN

In some cases, SSNs were required in the online applications. We assigned each fictitious applicant an SSN from the publicly available database of SSNs belonging to people deceased before 2007 (on National Archives and Records Administration (NARA) 2007). We manually selected different SSN files based on surname initials and varied starting digits to diversify the pool of SSNs assigned to the applicants. We also ruled out SSNs starting with 0. The steps requiring SSNs in the online applications are commonly hosted by third-party companies for verification purposes.

Questions

All online applications included some questions requiring responses from applicants. We only answered questions that were required and ignored all questions that were optional. To ensure the same answers across applicants, we asked the research assistants to fill in answers as the following: all daytime during the week (including weekends) for available date; a random day in the next two weeks from the application day for the start day; the average automotive sales salary ($35,000) for expected salary; having drivers licenses issued by their residential states (took "state-assigned" courses); not in any government subsidy programs such as SNAP; not wishing to disclose demographic information about gender, ethnicity, and veteran status; committing to only full-time roles; not having any certificate other than GED (if not high school diploma); having no felony history. Some application portals required at least one prior experience, and we used volunteer experience for candidates without prior experiences in such cases.

Application Details

To avoid suspicion and ensure the job listings are still open, we sent the four applications for each job at different times across three days. For example, if we sent the first application on Monday morning, each of the remaining three applications was sent during randomly selected blocks from the following eight: Monday afternoon, Monday evening, Tuesday morn-

ing, Tuesday afternoon, Tuesday evening, Wednesday morning, Wednesday afternoon, and Wednesday evening.

When applying, the research assistants also further checked for job requirements, whether the residential, prior employment (if any), and high school addresses matched the job location, and whether prior employment (if any) was at the same company as the job listing. Job listings with disqualifying requirements were discarded, and applications to jobs with disqualifying resumes were paused until all four resumes were ready to be sent. In the event of a job listing closing before we sent out all four applications, we used another job for the unsent conditions, such that all four treatment arms were applied across the two job listings.

C.1.4 Call Analysis

We considered first-time phone calls to each applicant from each company. We counted the phone calls for each candidate as follows. First, monitoring the dedicated phone line, we identified phone numbers associated with the companies we applied to by examining their displayed names (e.g., some numbers will show the name of the company) and their voicemail. This allowed us to create a dictionary of verified phone numbers and their associated companies. Second, we filtered phone calls from these verified numbers, and matched them to our fictitious applicants based on the recipient's (applicant's) phone number. In the cases where the same company called the same applicant multiple times, we only recorded the first call from this company. This resulted in a dataset of applicant-company pairs where at least one call was made. Lastly, we mapped all of the applicant-company pairs to the original dataset of applicants, associating the first-time calls with the treatments. We then summarized the results across the treatment arms to arrive at the final estimates of first-time calls for each treatment arm.

C.1.5 Local Market Thickness

For a group of companies, the general method for identifying automotive jobs within a municipality involved the following steps:

1. Filter all available job listings to include only full-time positions within a predetermined radius around the city (5 miles). If a company website did not allow search within a five-mile radius, we manually looked up listings in towns within a 5 mile radius.

2. Scrape job listing URLs and titles from each page of the filtered results. To avoid redundancy, only one URL was randomly selected for each unique job title.

3. Visit each selected URL to extract the text outlining the job requirements. Keywords like "required," "high school," "ability," "level," "experience," "diploma," or "degree"

were used to locate relevant information.

4. Assess job listings based on their education and experience requirements. Listings for managerial positions or those that required education beyond high school were disqualified.

5. Maintain a list of qualified job titles. If a newly scraped job listing had a title already present in the qualified list, it was automatically counted as qualified without further checks.

This general method required adjustments for certain companies due to variations in website structure or specific requirements. For instance, some companies' job listings might not have had a dedicated "requirements" section, requiring the use of different keywords and criteria to assess qualification. Other companies might have included remote work options or required bilingual skills, which needed to be excluded for the purpose of this analysis. The specific keywords and criteria used to identify qualified entry-level jobs were therefore tailored to each company to account for these variations.

Research assistants monitored the scraping process and randomly verified job titles. They also manually reviewed titles that were potentially not entry-level, such as "manager," and created lists of non-entry-level titles to exclude.

C.1.6 Callback Converstion Rate

To measure conversion rate from callbacks to employment, we recruited and surveyed 107 actual hiring managers through a recruiting agency, from automotive industry stores similar to the ones targeted in our study. We first asked about their job titles and their duration in these roles. Then, we proceeded to ask them about the conversion rate (from interview to actual job offer) in three blocks.

We first asked them for a "base conversion" rate by providing no (experience or racial) information about the candidate: "Suppose you have reviewed an applicant's resume for a job that requires minimal experience, such as a cashier or inventory clerk. You have already decided to invite him or her for an interview. In your experience, what are the chances that the applicant ends up being offered the job? Note that 50% chance means that the person is offered the job half of the time, and 100% chance means the applicant is offered the job every time."

We then showed them the experience treatment block, where they were asked to provide the conversion rates for two candidates: "Consider an applicant [with or without experience] who is interviewed. What is the chance that the applicant is offered a job after the interview?"

Lastly, we showed them the race treatment block, where we, similarly, asked about the conversion rates for another two candidates ("Consider a [Black or White] applicant who is interviewed. What is the chance that the applicant is offered a job after the interview?").

The order of the two candidates was randomized within each block. Afterwards, the surveyed managers answered the demographic questions and one open-text question for them to leave comments.

C.2 Experimental IA

C.2.1 Preparing Letter Content

Focusing on STEM majors, we chose "Mechanical Engineering" and "Computer and Information Science" as the majors of the candidates. To avoid potential confounding effects of school rank and private universities, we chose large public universities similar in major rankings for each candidate: Penn State University for Mechanical Engineering and Ohio State University for Computer and Information Science. Each candidate had three prior experiences in their major-related field, where the job titles and descriptions were generated by ChatGPT and manually reviewed. We also used the InfoGroup 2022 business data to find their prior employers. We first filtered the data based on location and primary SIC code (indicating industry) to only preserve companies whose industries are related to these two majors, and which are located in cities near the universities (Columbus, OH and Philadelphia, PA). We then used the first three companies for each city as the prior employers. Since the InfoGroup dataset does not seem to be sorted based on any variables, selecting the first three companies was not suggestive of any attributes.

The durations of employment were determined randomly. For the most recent employment, we subtracted from the day of resume creation a random integer from uniform $[7, 28]$ as the most recent employment end date. This suggests all candidates had been unemployed for at most four weeks at the time of resume creation. For the second most recent employment end date, we further subtracted from the most recent end date a randomly selected number from uniform $[0.8, 1.2]$ times 365. This indicates that the time between the second and the most recent end dates were between 80% of a year and 120% of a year. For the earliest employment, we repeated the same process but subtracted from the second most recent end date. For the two most recent employment start dates, we added a randomly selected integer from uniform $[14, 45]$ to the previous end dates, indicating that the gap between employments was between 14 and 45 days. For the earliest employment, we determined the start date by adding a random integer from uniform $[1, 60]$ to the same day of the month as the employment end date, but one year earlier and in July. For example, if the employment ended on Jan 15, 2024, we added the random integer to July 15, 2023. This is because the earliest employment should be the candidates' first employment after university, and our

method would indicate this employment happened in the summer or fall of the graduation year. We then assumed all candidates graduated in early June of the same year as their earliest employment. The university start date was determined by first subtracting the university end dates by four years, and then adding a random integer from uniform $[60, 100]$ such that university started in August or September.

For names of the candidates, we chose two male first names and two female first names. The four first names were mapped to four distinct common surnames. The first names are not suggestive of any minority race, and surnames all commonly belong to White people. The surnames were randomly selected from Kline et al. (2024) (*Bauer, Mast, Hostetler, Hershberger*), and the first names were chosen from the top four common names by gender during the 2000s listed by SSA to match the age of the candidates (*Emily, Olivia, Joshua, Matthew*). We grouped the names into two pairs, where each pair consisted of one male and one female name. Within each pair, the candidates shared identical education and job experience. The two pairs took up the two majors, corresponding universities, and related experiences, respectively.

C.2.2 Generating Example Letters

With this information, we prompted ChatGPT to write recommendation letters by asking for a 300-word letter using all the education and prior experience information for each candidate. All such information was provided in bullet-points, similar to a resume layout. After generating the set of four letters, we checked for both word sentiment and word categories according to the LIWC dictionary (Kaplan et al. 2024). We verified that the letters exhibited similar levels of bias in lexical content across the dimensions examined in Wan et al. (2023): letters written for male candidates were significantly more formal, positive, and agentic than those written for female candidates.

C.2.3 Generating Multiple Letters

Once we generated the initial four letters, we changed the ChatGPT prompt to request additional letters, similar in terms of sentiment and type of words, for the same candidates and the same positions outlined in the previous subsection. In this manner, we created 25 letters for each candidate. These correspond to the Endogenous-m and Endogenous-f letters for the male and female candidates, respectively. Next, we took 50 letters from the Endogenous-m set and replaced the names with their female counterparts', and the pronouns with she/her pronouns. These corresponded to the Exogenous-f letters. We used the same process to create the Exogenous-m letters.

E-lit-009: Reasonable Doubt: Experimental Detection of Job-Level Employment Discrimination

Patrick Kline and Christopher Walters | Econometrica 89(2): 765-792 (NBER Working Paper 26861) | 2021-03

https://eml.berkeley.edu/~pkline/papers/Kline_Walters_RD_2020.pdf

About this source

The paper's central result is that job-level discrimination is extremely concentrated, so the market average conceals rather than summarizes employer behavior. The lower bound on the share of jobs with different white and black callback probabilities is at least 13 percent in Bertrand and Mullainathan and at least 17 percent in Nunley et al.; a fitted two-type mixed logit independently puts the discriminating share at about 17 percent, close to the non-parametric bound. Among those firms the behavior is not marginal: the estimated odds of a callback are roughly 53 times higher for a white than a black application, which is far more extreme than the 2.6 to 3.1 percentage point pooled gaps the original studies reported. The job-level standard deviation of the black-white gap (0.082 in Bertrand and Mullainathan) is more than twice its mean, and higher moments show left skew and excess kurtosis - a long tail of heavy discriminators. This heterogeneity has forensic value rather than just being noise: at least 72 percent of Bertrand and Mullainathan jobs that called back two whites and neither black applicant were discriminating, and at least 85 percent of such jobs in Nunley et al. (90 percent in a three-white/one-black design). The Mexico City data show discrimination running in both directions at once - at least 14 percent of jobs against men and at least 6 percent against women - so the small average gap in favor of women masks offsetting employer-level behavior. Finally, the authors show the design implication: with ten applications per job a regulator using an 80 percent posterior threshold could detect about 7.5 percent of discriminating employers while investigating under 0.2 percent of innocent ones, whereas the standard four-application design detects almost none. The paper also rebuts one common reading of the data, showing that the apparent correlation between the size of the gap and the employer's overall callback rate is a mechanical boundary effect, since the propensity to discriminate is nearly independent of the white callback rate.

Bound as the original PDF. Retrieved 2026-09-10. sha256 bead0c95df3b69e7...

Bound into an Economic Briefing · by Claude, may contain mistakes

Full document begins on the next page.

E-lit-009: https://eml.berkeley.edu/~pkline/papers/Kline_Walters_RD_2020.pdf

NBER WORKING PAPER SERIES

REASONABLE DOUBT:
EXPERIMENTAL DETECTION OF JOB-LEVEL EMPLOYMENT DISCRIMINATION

Patrick M. Kline
Christopher R. Walters

Working Paper 26861
http://www.nber.org/papers/w26861

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2020, Revised August 2020

This paper previously circulated under the title “Audits as Evidence: Experiments, Ensembles, and Enforcement.” We thank Isaiah Andrews, Tim Armstrong, Kerwin Charles, Sendhil Mullainathan, and Andres Santos for helpful conversations related to this project, and Eva Arceo-Gomez, Ray Campos-Vasquez, and John Nunley for providing data. We also thank participants at the UC Berkeley labor and econometrics seminars, the Y-Rise External Validity conference, the 2019 NBER Summer Institute, University of British Columbia, UC Irvine, Harris School of Public Policy, the Tinbergen Institute, the Stockholm School of Economics, the University of Oslo, the 2019 All California Economics Conference, the University of Michigan, Stanford University, Harvard University, and Clemson University for useful feedback. Evan Rose and Benjamin Scuderi provided outstanding research assistance. This project was supported by a Russell Sage Foundation Presidential grant. 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.

© 2020 by Patrick M. Kline and Christopher R. Walters. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Reasonable Doubt: Experimental Detection of Job-Level Employment Discrimination
Patrick M. Kline and Christopher R. Walters
NBER Working Paper No. 26861
March 2020, Revised August 2020
JEL No. C14,C44,C9,J7,J71,K31,K42

ABSTRACT

This paper develops methods for detecting discrimination by individual employers using correspondence experiments that send fictitious resumes to real job openings. We establish identification of higher moments of the distribution of job-level callback rates as a function of the number of resumes sent to each job and propose shape-constrained estimators of these moments. Applying our methods to three experimental datasets, we find striking job-level heterogeneity in the extent to which callback probabilities differ by race or sex. Estimates of higher moments reveal that while most jobs barely discriminate, a few discriminate heavily. These moment estimates are then used to bound the share of jobs that discriminate and the posterior probability that each individual job is engaged in discrimination. In a recent experiment manipulating racially distinctive names, we find that at least 85% of jobs that contact both of two white applications and neither of two black applications are engaged in discrimination. To assess the potential value of our methods for regulators, we consider the accuracy of decision rules for investigating suspicious callback behavior in various experimental designs under a simple two-type model that rationalizes the experimental data. Though we estimate that only 17% of employers discriminate on the basis of race, we find that an experiment sending 10 applications to each job would enable detection of 7-10% of discriminatory jobs while yielding Type I error rates below 0.2%. A minimax decision rule acknowledging partial identification of the distribution of callback rates yields only slightly fewer investigations than a Bayes decision rule based on the two-type model. These findings suggest illegal labor market discrimination can be reliably monitored with relatively small modifications to existing correspondence designs.

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

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

It is illegal to use information on race, sex, or age to make employment decisions in the United States.[1] The voluminous empirical literature on labor market discrimination has focused primarily on establishing whether *markets* discriminate against particular groups of workers on average (Altonji and Blank, 1999; Guryan and Charles, 2013). However, a finding of market-level discrimination provides little guidance to regulators tasked with enforcing anti-discrimination law, who must decide which specific employers to investigate (US Equal Employment Opportunity Commission, 2016). Indeed, classic models of discrimination emphasize that market-level outcomes provide limited guidance regarding the underlying *distribution* of discrimination across employers (Becker, 1957). This paper extends the frontier by developing new tools to characterize discriminatory behavior in markets and detect discrimination by individual employers.

Our approach adapts insights from the literature on empirical Bayes (EB) analysis of large scale testing problems (Efron, 2012) to the study of correspondence experiments that submit fictitious applications with randomly generated characteristics to actual job vacancies (Bertrand and Duflo, 2017 provide a review). Since the influential work of Bertrand and Mullainathan (2004), correspondence experiments typically sample thousands of jobs and send each a few applications with distinctive names that signify race or sex. Our basic insight is that such studies are best viewed as *ensembles* of exchangeable micro-experiments. From the ensemble, one can infer properties of the distribution of discriminatory behavior which can, in turn, be used to form empirical posteriors about the probability that any given job is discriminating.

As in classic EB analyses of count data (Efron and Morris, 1975; Brown, 2008), we treat callback outcomes as independent Bernoulli trials governed by job- and race (or sex)-specific callback probabilities, a modeling choice we show closely approximates callback behavior in correspondence experiments. Because few applications are sent to each job, the distribution of job-specific callback probabilities is under-identified, invalidating standard non-parametric EB approaches (e.g., Efron, 2016). However, we establish identification of a set of moments of the joint distribution of white and black callback probabilities determined by the number of applications sent to each job. To estimate these moments, we propose a Shape-Constrained Generalized Method of Moments (SCGMM) estimator that requires the estimated moments to be consistent with a proper bivariate probability distribution. Applying this estimator to three experimental datasets reveals tremendous heterogeneity across jobs in the extent to which callback probabilities differ by race or sex. Estimates of third and higher moments reveal that while most jobs barely discriminate, a few discriminate heavily.

Extending classic results on the identification of False Discovery Rates (Benjamini and Hochberg, 1995; Efron et al., 2001; Storey, 2002), we compute sharp lower bounds on the share of jobs engaged in discrimination given the identified moments. In the Bertrand and Mullanaithan experiment, we estimate that at least 13% of jobs discriminate against black applicants. The corresponding estimate in a more recent study by Nunley et al. (2015) is 17%. In a study by Arceo-Gomez and Campos-

[1] Title VII of the Civil Rights Act of 1964 prohibits employment discrimination on the basis of race and sex, while the Age Discrimination Act of 1975 prohibits certain forms of discrimination on the basis of age.

Vasquez (2014), we find that at least 6% of jobs discriminate against women, while at least 14% discriminate against men. These population shares are then used to compute bounds on posterior probabilities that particular jobs are discriminating given their callback patterns. We find that these posterior bounds are often highly informative even with few applications per job. For example, we estimate that at least 72% of jobs calling both of two white applicants and neither of two black applicants in the Bertrand and Mullainathan (2004) experiment are engaged in discrimination, while at least 85% of such jobs in the Nunley et al. (2015) experiment are discriminating. In the Arceo-Gomez and Campos-Vasquez (2014) experiment we find that at least 97% of the jobs that call back four women and no men discriminate against men, while at least 88% of jobs that call back four men and no women discriminate against women.

To explore the potential policy implications of our findings, we assess the prospects for systematically detecting discriminators based on callback evidence generated by alternative hypothetical correspondence study designs. This investigation is based on detection/error tradeoffs that arise under a parametric two-type model fit to the Nunley et al. (2015) data. With only two white and two black applications per job, it is difficult to reliably identify discriminating employers. But with 10 applications per job, we find that a regulator who knows the joint distribution of callback probabilities can correctly identify 7% of discriminating jobs while incurring Type I error rates of less than 0.2%.

Finally, to probe the sensitivity of these conclusions to our modeling assumptions, we consider the problem of a hypothetical regulator who knows only the identified moments of the distribution of callback probabilities. This regulator decides which jobs to investigate using a minimax decision rule that minimizes the maximum risk consistent with the known moments. We develop a tractable approach to estimating the maximum risk function and find that a minimax regulator investigates only slightly fewer jobs than would a Bayesian regulator who knows the joint distribution of callback probabilities. This robustness emerges because the risk function is nearly identified at realistic posterior thresholds that might be used to trigger investigations.

Our results illustrate the potential of experimental methods to assist with regulatory enforcement of anti-discrimination laws. Because employers vary tremendously in their propensity to discriminate against protected groups, regulators face a difficult inferential task. Our findings suggest correspondence experiments can be paired with simple decision rules to reliably identify discriminators. More generally, the methods developed here provide a tractable empirical framework for making decisions when the population distribution of unit heterogeneity is only partially identified by an experiment. Candidate applications of these methods include targeting of workplace safety inspections based on experimental audit data (Levine et al., 2012), detecting violations of rationality with laboratory choice experiments (Halevy et al., 2018), and evaluating the performance of individual teachers and schools based upon student achievement data (Chetty et al., 2014a; Angrist et al., 2017).

2 Defining Discrimination

We now develop a formal notion of discrimination tailored to the analysis of correspondence experiments. To simplify exposition we focus on race, which we code as binary ("white" / "black"). Suppose that we have a sample of J jobs with active vacancies. To each of these jobs, we send L_w applications with distinctively white names and L_b applications with distinctively black names as in Bertrand and Mullainathan (2004), for a total of $L = L_w + L_b$ applications. Denote the race associated with the name used in application $\ell \in \{1, ..., L\}$ to job $j \in \{1, ..., J\}$ as $R_{j\ell} \in \{w, b\}$. The potential callback function $Y_{j\ell} : \{w, b\} \to \{0, 1\}$ indicates whether job j would call back application ℓ as a function of that applicant's assigned race. Observed callbacks are then given by $Y_{j\ell} = Y_{j\ell}(R_{j\ell})$.

When $Y_{j\ell}(w) \neq Y_{j\ell}(b)$ job j has engaged in racial discrimination with application ℓ. Notably, even if racially distinctive names influence employer behavior only through their role as a proxy for parental background (Fryer and Levitt, 2004), using the names at any point in the hiring process is likely to be viewed by courts as a pretext for discrimination.[2] While courts are typically interested in establishing whether a particular plaintiff experienced discrimination in precisely this sense, we will take the perspective of a regulator tasked with assessing prospectively whether an employer systematically treats applicants differently based upon race. For example, the mission of the US Equal Employment Opportunity Commission (EEOC) is to "***prevent*** and remedy unlawful employment discrimination and advance equal opportunity for all in the workplace" (emphasis added). The following assumption formalizes this prospective notion of discrimination at the employer level.

Assumption 1. *Callbacks are race- and job-specific Bernoulli trials:*

$$Y_{j\ell}(r) \,|R_{j1}...R_{jL} \overset{iid}{\sim} Bernoulli(p_{jr}) \quad for\ r \in \{w, b\}.$$

Note that random assignment of racially distinctive names to applications guarantees independence of $Y_{j\ell}(r)$ from $\{R_{jk}\}_{k=1}^{L}$. The key behavioral restriction in Assumption 1 is that the $\{Y_{j\ell}(r)\}_{\ell=1}^{L}$ are *iid*, which rules out, for example, scenarios in which a job calls back the first qualified applicant and disregards all subsequent applications.[3] We discuss below how to test for such violations. The probability p_{jr} may be interpreted as the callback rate that would emerge in a hypothetical experiment in which a large number of applications of race r are sent to job j.[4]

Letting $C_{jr} = \sum_{\ell=1}^{L} 1\{R_{j\ell} = r\} Y_{j\ell}$ denote the number of applications of race r to job j that were called back, Assumption 1 implies the probability $\Pr(C_{jw} = c_w, C_{jb} = c_b | p_{jw}, p_{jb})$ that employer j

[2] See, e.g., the discussion in U.S. Equal Employment Opportunity Commission v. Target Corporation, 460 F.3d 946, 7th Cir. Wis. 2006 and footnote 27 of Fryer and Levitt (2004).

[3] One could equivalently view such behavior as a violation of our specification of potential outcomes, which builds in the Stable Unit Treatment Value Assumption of Rubin (1980).

[4] If hundreds of applications were sent to a single job the employer would likely be overwhelmed and Assumption 1 would fail. We show below, however, that this assumption provides a suitable approximation to an experiment with 8 applications, which is an unusually large choice of L.

calls back c_w white applications and c_b black applications is:

$$f\left(c_w, c_b | p_{jw}, p_{jb}\right) \quad = \quad \binom{L_w}{c_w}\binom{L_b}{c_b} p_{jw}^{c_w}\left(1-p_{jw}\right)^{L_w-c_w} p_{jb}^{c_b}\left(1-p_{jb}\right)^{L_b-c_b}. \tag{1}$$

We are now ready to offer a job-level definition of discrimination, which we will henceforth refer to simply as discrimination.

Definition. *Job j engages in discrimination when $p_{jb} \neq p_{jw}$.*

Discriminatory jobs are labeled with the indicator function $D_j = 1\{p_{jb} \neq p_{jw}\}$. This definition is prospective in that an employer with $D_j = 1$ will eventually discriminate against an applicant even if it has not done so yet.

3 Ensembles and Posteriors

The above framework treats each job's callback decisions as a set of race-specific Bernoulli trials. We next consider what can be learned from an ensemble of experiments conducted at many jobs. This idea is formalized in the following exchangeability assumption on the jobs.

Assumption 2. *Race-specific callback probabilities are independent and identically distributed:*

$$p_{jw}, p_{jb} \overset{iid}{\sim} G\left(\cdot,\cdot\right).$$

The distribution function $G : [0,1]^2 \to [0,1]$ describes the population of jobs from which a study samples. In practice, audit studies usually draw small random samples of jobs from online job boards. The *iid* assumption abstracts from the fact that there are a finite number of jobs on these boards. Note that by virtue of random assignment p_{jw} and p_{jb} are independent of the racial mix of applications to job j as well as any other resume characteristics that are randomized.

Assumption 2 implies that the unconditional distribution of callbacks can be expressed as a mixture of binomial trials. We denote the unconditional probability of observing the callback vector (c_w, c_b) by

$$\bar{f}\left(c_w, c_b\right) \quad = \quad \int f\left(c_w, c_b | p_w, p_b\right) dG\left(p_w, p_b\right). \tag{2}$$

The distribution G will serve as a key object of interest in our analysis. One reason for interest in G is that it characterizes both the prevalence and severity of discrimination in a population. Prevalence is captured by the proportion of jobs that are engaged in discrimination, which can be written:

$$\bar{\pi} = \Pr\left(D_j = 1\right) = \int_{p_w \neq p_b} dG\left(p_w, p_b\right).$$

Likewise, the severity of discrimination can be summarized by moments of the form $\int \left(p_w - p_b\right)^m dG\left(p_w, p_b\right)$, which will equal zero for any $m \in \mathbb{N}$ if all jobs call back applicants independently of race.

A second reason for interest in G lies in its potential forensic value as a tool for identifying which jobs are discriminating. The quantity $\pi(c_w, c_b) = \Pr(D_j = 1 | C_{jw} = c_w, C_{jb} = c_b)$ gives the prevalence of discrimination among jobs with callback vector (c_w, c_b). We can also think of $\pi(C_{jw}, C_{jb})$ as the posterior probability that a job is discriminating given the callback evidence (C_{jw}, C_{jb}). Invoking Bayes' rule, $\pi(c_w, c_b)$ can be expressed as a functional of the "prior" distribution G:

$$\begin{aligned}
\pi(c_w, c_b) &= \frac{\Pr(C_{jw} = c_w, C_{jb} = c_b | D_j = 1)\,\bar{\pi}}{\bar{f}(c_w, c_b)} \\
&= \frac{\int_{p_w \neq p_b} f(c_w, c_b | p_w, p_b)\, dG(p_w, p_b)}{\bar{f}(c_w, c_b)} \\
&= \mathcal{P}\left(\underbrace{c_w, c_b}_{\text{direct}}, \underbrace{G}_{\text{indirect}}\right).
\end{aligned}$$

The dependence of $\pi(c_w, c_b)$ on G is an example of what Efron (2010) refers to as "indirect evidence." To understand the logic of incorporating indirect evidence, suppose $\bar{\pi} = 0$ so that no jobs discriminate. Then $\pi(C_{jw}, C_{jb}) = 0$ with probability one – any seemingly suspicious callback decisions are due to chance. Likewise, if $\bar{\pi} = 1$, all jobs are discriminators and there is no need for direct evidence on the behavior of particular jobs. But in intermediate cases, where some share of jobs are discriminators, and some are not, it is rational to blend the direct evidence from a particular job with contextual information on the population from which that job was drawn.

Empirical Bayes approaches seek to form empirical posteriors $\mathcal{P}\left(c_w, c_b, \hat{G}\right)$ that substitute the unknown G with an estimator $\hat{G}$. Important applications of this idea arise in the literature on multiple hypothesis testing, where a key concept is the False Discovery Rate (Benjamini and Hochberg, 1995), which can be thought of as a posterior estimate of the probability that a given null hypothesis is true (Efron et al., 2001; Storey, 2002, 2003). In our setting, the False Discovery Rate corresponds to the proportion $1 - \pi(c_w, c_b)$ of jobs with evidence vector (c_w, c_b) that are not discriminating. Appendix A develops this connection in more detail.

4 Identification of G

Each job's realized callback rates $(C_{jw}/L_w, C_{jb}/L_b)$ provide noisy estimates of the latent callback probabilities (p_{jw}, p_{jb}). The binomial structure of this noise is not classical which leads point identification of G to fail when the number of applications per job is small.[5] In this section we establish that certain moments of G are nonetheless identified by simple linear transformations of unconditional callback probabilities. We then proceed to derive bounds on the posterior probability

[5] If L were to grow large, one could invoke a normal approximation on each job's sample callback rates and then apply a variance stabilizing transform to make the noise approximately homoscedastic, as in classic EB studies of batting averages (Efron and Morris, 1975; Brown, 2008). With homoscedastic normal estimation error, G could then be estimated via deconvolution (e.g., as in Efron, 2016). However, Brown (2008) cautions against using such approximations with 10 or fewer observations per group.

function $\pi(c_w, c_b)$ consistent with these moments.

Moments

From (2) we can write

$$\begin{aligned}\bar{f}(c_w, c_b) &= \binom{L_w}{c_w}\binom{L_b}{c_b}\mathbb{E}\left[p_{jw}^{c_w}(1-p_{jw})^{L_w-c_w}\, p_{jb}^{c_b}(1-p_{jb})^{L_b-c_b}\right] \\ &= \binom{L_w}{c_w}\binom{L_b}{c_b}\sum_{m=0}^{L_w-c_w}\sum_{n=0}^{L_b-c_w}(-1)^{m+n}\binom{L_w-c_w}{m}\binom{L_b-c_b}{n}\mathbb{E}\left[p_{jw}^{c_w+m}p_{jb}^{c_b+n}\right], \quad (3)\end{aligned}$$

where $\mathbb{E}[\cdot]$ denotes the expectation with respect to G. Hence, the reduced form callback rates can be written as linear functions of uncentered moments of the latent callback probabilities.

We index these moments via the function:

$$\mu(m,n) \equiv \mathbb{E}\left[p_{jw}^m p_{jb}^n\right] \quad \text{for } (m,n) \in \mathbb{N}^2.$$

Letting $\bar{f} = \left(\bar{f}(1,0), ..., \bar{f}(L_w,0), ..., \bar{f}(L_w,L_b)\right)'$ denote the vector of frequencies for all possible callback outcomes excluding $(0,0)$ and $\mu = (\mu(1,0), ..., \mu(L_w,0), ..., \mu(L_w,L_b))'$ the corresponding list of moments, we can write the equations in (3) as a linear system $\bar{f} = B\mu$, where B is a known non-singular square matrix of binomial coefficients. Inverting the linear system yields

$$\mu = B^{-1}\bar{f}, \quad (4)$$

which immediately implies the following Lemma.

Lemma 1 (Identification of Moments)**.** *Under Assumptions 1 and 2 and for a given application design (L_w, L_b), all moments $\mu(m,n)$ for $0 \le m \le L_w$ and $0 \le n \le L_b$ are identified.*

Lemma 1 formalizes the sense in which the identifying power of a correspondence experiment is increasing in the number of applications of each type sent to each job. As $\min\{L_w, L_b\}$ grows large, the entire joint distribution of black and white callback probabilities becomes identified, which motivates non-parametric deconvolution approaches such as Efron (2016). With a finite number of applications per job, Lemma 1 establishes which moments of G are identified from (4).

For example, when $L_w = L_b = 2$ as in Bertrand and Mullainathan (2004), Lemma 1 establishes identification of the first two moments of black and white callback probabilities. From these moments, one can identify the following measure of job-level heterogeneity in discriminatory behavior:

$$\mathbb{V}[p_{jb} - p_{jw}] = \mu(0,2) + \mu(2,0) - 2\mu(1,1) - \mu(0,1)^2 - \mu(1,0)^2 + 2\mu(0,1)\,\mu(1,0).$$

With more applications per job, higher moments of the distribution of job level discrimination become identified. Furthermore, when the application design (L_w, L_b) varies randomly across jobs, some moments of G will be over-identified. We later exploit these over-identifying restrictions in estimation to improve precision and test our modeling assumptions.

Analytic Bound on Posterior Probabilities

Though the study of moments of the callback distribution G can shed light on underlying heterogeneity in callback behavior, the posterior probability $\pi(c_w, c_b)$ need not admit a representation in terms of a finite number of moments. However, a simple analytic bound on the posterior can be derived from an application of Bayes' rule that conditions on the total number of callbacks $C_{jb}+C_{jw}$ to job j. Let $\bar{f}_t(c_w) = \Pr(C_{jw} = c_w, C_{jb} = t - c_w | C_{jb} + C_{jw} = t)$ denote the probability mass function for white callbacks in the stratum of jobs that call back t applicants in total, and let $\bar{f}_t^0(c_w) = \binom{L_w}{c_w}\binom{L_b}{t-c_w}/\binom{L}{t}$ denote the corresponding probability mass function that would arise under Assumptions 1 and 2 if no discrimination were present in this stratum. Finally, let $\bar{\pi}_t = \Pr(D_j = 1 | C_{jb} + C_{jw} = t)$ denote the share of jobs calling back t total applicants that are engaged in discrimination.

One can write the posterior probability of discrimination in terms of these objects as follows:

$$\pi(c_w, c_b) = 1 - \underbrace{(1 - \bar{\pi}_{c_w+c_b})}_{\text{prior that } D=0} \underbrace{\bar{f}^0_{c_w+c_b}(c_w)}_{\text{likelihood if } D=0} / \underbrace{\bar{f}_{c_w+c_b}(c_w)}_{\text{marginal likelihood}}.$$

Because the probability mass function $\bar{f}_t(c_w)$ is directly identified by the experiment, the only nuisance parameters entering the posterior are the stratum specific priors $\{\bar{\pi}_t\}_{t=0}^L$. For example, in an experiment with $L_w = L_b = 2$, the prevalence of discrimination among jobs that call back only white applications is $\pi(2,0) = 1 - (1 - \bar{\pi}_2)(1/6)/\bar{f}_2(2)$. By contrast, in this design, the exact p-value on the null hypothesis that a job calling back only white applications is *not* discriminating is simply $\bar{f}_2^0(2) = 1/6$. Note that $\pi(2,0) > \bar{f}_2^0(2)$ if and only if $\bar{\pi}_2 > 1 - 5\bar{f}_2(2)$, which highlights the crucial role of the stratum prior in drawing posterior inferences.

The following Lemma, which is proved in Appendix B, provides a tractable bound on both the stratum prior $\bar{\pi}_t$ and, consequently, the posterior prevalence function $\pi(c_w, t - c_w)$.

Lemma 2 (Bounds on Stratum Prior and Posterior)**.**

$$i)\ \bar{\pi}_t \geq \max_{c_w \in \{0,...,t\}} \max\left\{\frac{\bar{f}_t^0(c_w) - \bar{f}_t(c_w)}{\bar{f}_t^0(c_w)}, \frac{\bar{f}_t(c_w) - \bar{f}_t^0(c_w)}{1 - \bar{f}_t^0(c_w)}\right\},\ t \in \{1, ..., L-1\},$$

$$ii)\ \pi(c_w, t - c_w) \geq 1 - \frac{\bar{f}_t^0(c_w)}{\bar{f}_t(c_w)} \min_{c'_w \in \{0,...,t\}} \min\left\{\frac{\bar{f}_t(c'_w)}{\bar{f}_t^0(c'_w)}, \frac{1 - \bar{f}_t(c'_w)}{1 - \bar{f}_t^0(c'_w)}\right\},\ t \in \{c_w, ..., L-1\}.$$

Part i) of this Lemma shows that the experiment places a lower bound on the share of jobs engaged in discrimination in each callback stratum that is increasing in the discrepancy between the distribution of callback outcomes and the distribution predicted by the non-discrimination null. Efron et al. (2001, p. 1154) proposed an analogous bound to control the False Discovery Rate in a multiple testing analysis of a microarray experiment. Part ii) establishes via standard Bayesian updating arguments that the bound on the prior translates into a corresponding lower bound on the posterior.

Sharp Bounds

While the bounds in Lemma 2 are easy to compute, they need not be sharp, as restrictions across callback strata have been ignored. A lower bound on the stratum prior $\bar{\pi}_t$ that exploits all of the logical restrictions in our framework can be written as the solution to the following constrained optimization problem:

$$\min_{G \in \mathscr{G}} 1 - \frac{\binom{L}{t}}{\sum_{c'_w=0}^{t} \bar{f}\left(c'_w, t - c'_w\right)} \int p^t \left(1-p\right)^{L-t} dG\left(p,p\right), \tag{5}$$

$$s.t.\ \ \bar{f}\left(c_w, c_b\right) = \binom{L_w}{c_w}\binom{L_b}{c_b} \int p_w^{c_w} \left(1-p_w\right)^{L_w - c_w} p_b^{c_b} \left(1-p_b\right)^{L_b - c_b} dG\left(p_w, p_b\right), \tag{6}$$
$$\text{for}\left(c_w = 0, .., L_w; c_b = 0, .., L_b\right).$$

To make this problem computationally tractable, we consider a space $\mathscr{G}$ of discretized approximations to the unknown distribution function G.[6] Because both the objective and constraints are linear in the probability mass function associated with G, we can apply linear programming (LP) routines to compute bounds given an estimate of the callback probabilities $\left\{\bar{f}\left(c_w, c_b\right)\right\}_{c_w, c_b}$. Details of our computational procedure are given in Appendix C. Since the distribution G is not indexed by t, the solution to (5) enforces constraints across callback strata. Consequently, we may obtain informative bounds on the share of discriminatory jobs even among those jobs that call no (or all) applications back.[7]

Analogous LP formulations can be used to bound any linear functional of G. For example, the lower bound on the unconditional prevalence of discrimination $\bar{\pi}$ is obtained by replacing the objective in (5) with $\min_{G(\cdot,\cdot)\in\mathscr{G}} 1 - \int dG\left(p,p\right)$. Likewise, we can bound from below the proportion of jobs discriminating against whites by replacing the objective in (5) with $\min_{G(\cdot,\cdot)\in\mathscr{G}} \int_{p_w<p_b} dG\left(p_w, p_b\right)$. We leverage these insights to bound from below the prevalence of a variety of directional notions of discrimination in the empirical work to follow.

5 Data

We apply our methods to data from three correspondence experiments summarized in Table I. Bertrand and Mullainathan (BM, 2004) applied to 1,112 job openings in Boston and Chicago, submitting four applications to each job. Of the four applications, two were assigned black-sounding names while the remaining two were assigned white-sounding names. The callback rate to applications with black sounding names was 3.1 percentage points lower than to applications with white sounding names.

[6] See Noubiap et al. (2001) for a closely related approach and an asymptotic analysis of the effects of discretization.

[7] Suppose, for instance, that G is a two type mixture with $\Pr\left(p_{jw}=1, p_{jb}=1\right) = 1/2$ and $\Pr\left(p_{jw}=1/2, p_{jb}=0\right) = 1/2$. Then all jobs with zero callbacks must be discriminators.

Table I: Descriptive statistics for resume correspondance studies

	Bertrand & Mullainathan (1)	Nunley et al. (2)	Arceo-Gomez & Campos-Vasquez (3)
Number of jobs	1,112	2,305	799
Applications per job	4	4	8
Treatment/control	Black/white	Black/white	Male/female
Callback rates: Total	0.079	0.167	0.123
Treatment	0.063	0.154	0.108
Control	0.094	0.180	0.138
Difference	-0.031	-0.026	-0.033
	(0.007)	(0.007)	(0.008)

Notes: This table reports sample characteristics based on data from three resume correspondence experiments. Columns (1) and (2) show statistics from Bertrand and Mullainathan's (2004) and Nunley et al.'s (2015) studies of racial discrimination in the United States. Column (3) reports statistics from Arceo-Gomez & Campos-Vasquez's (2014) study of gender discrimination in Mexico. Standard errors for treatment/control differences, clustered at the job level, are in parentheses.

Nunley et al. (NPRS, 2015) studied racial discrimination in the market for new college graduates by applying to 2,305 listings on an online job board, again sending four resumes per job opening. Unlike BM, the names assigned to the four resumes were sampled without replacement from a pool of eight names, four of which were distinctively black and four of which were distinctively white. This led the share of black names sent to each job to vary randomly in increments of 25% from 0% to 100%. The overall callback rate in the NPRS study was more than twice as high as in the BM study, perhaps because the fictitious applicants were more highly educated. On average, black names had a 2.6 percentage point lower callback rate than white names.

Arceo-Gomez and Campos-Vasquez (AGCV, 2014) applied to 799 job openings through an online job portal in a study of race and gender discrimination in Mexico City, Mexico. AGCV sent eight fictitious applications to each job, and the applicants were all recent college graduates. While the AGCV experiment looks at a different context than BM or NPRS, this data set allows us to demonstrate the gains from doubling the number of applications per job opening. To illustrate the identifying power of sending four applications to each protected group, we focus on gender in this experiment, as AGCV used a three-category definition of race.[8] In the AGCV experiment, women were 3.3 percentage points more likely to receive callbacks than men.

[8] In principle the methods developed here could be extended to a multivariate distribution of callback probabilities for three or more groups. Section 8 explores this possibility by allowing callback rates to differ by resume quality in addition to race or gender.

6 Are Callbacks Independent Trials?

We begin by considering tests of the Bernoulli trials assumption that undergirds our econometric framework. This assumption would be violated if the likelihood of a callback depends not just on an application's own characteristics but also on the characteristics of other applications sent to the same job. To assess this possibility, we fit linear probability models of the form:

$$Y_{j\ell} = \lambda_0 + X'_{j\ell}\lambda_1 + \bar{X}'_{j\ell}\lambda_2 + \varepsilon_{j\ell}, \tag{7}$$

where $X_{j\ell}$ is a vector of application characteristics and $\bar{X}_{j\ell} = (L-1)^{-1}\sum_{k\neq\ell} X_{jk}$ gives the "leave out" mean of those characteristics among the applications sent to job j excluding application ℓ. While the coefficient vector λ_1 gives the direct effect of application characteristics on callbacks, the coefficient vector λ_2 captures the "peer effect" of other applications to the same job on application ℓ's callback propensity. Assumption 1 restricts these peer effects to be zero ($\lambda_2 = 0$).

For OLS estimates of λ_2 to identify causal effects, $\bar{X}_{j\ell}$ must be uncorrelated with any omitted application characteristics $Z_{j\ell}$ that influence callbacks. We therefore focus on the NPRS study which assigned both race and a large number of other application characteristics independently of each other and across applications.[9]

Columns 1 and 2 of Table II report estimates of the parameters in (7) for the NPRS study, with each row showing the coefficients from a separate regression. While applications with distinctively black names are significantly less likely to be called back, we find no significant effect on callback probabilities of changing the racial mix of the other 3 applications to the same job. Across the 12 covariates we consider only one (an indicator for 3+ months of unemployment) finds a significant peer effect at conventional levels, and a joint test fails to reject that all of the leave out mean coefficients are zero ($p = 0.45$). As another composite test, we report the results of a model in which the peer effects are restricted to be proportional to the main effects of the application's own characteristics $X_{j\ell}$. The row titled "predicted callback rate" pools all the application characteristics into an index $X'_{j\ell}\hat{\lambda}_{1(j)}$ where $\hat{\lambda}_{1(j)}$ is the leave out OLS coefficient vector obtained from regressing the callback indicator on application covariates after leaving out all applications to job j. A unit increase in $X'_{j\ell}\hat{\lambda}_{1(j)}$ is associated with roughly half of a callback on average. Though $X'_{j\ell}\hat{\lambda}_{1(j)}$ strongly predicts callbacks, its average value among competing applications $(L-1)^{-1}\sum_{k\neq\ell} X'_{jk}\hat{\lambda}_{1(j)}$ has no statistically discernible impact on callbacks.

[9] In contrast, BM assigned application characteristics according to their joint distribution in a training sample, making it likely that the characteristics we study $X_{j\ell}$ are correlated with other omitted characteristics $Z_{j\ell}$ that predict callbacks. The application characteristics were also chosen to yield a good match with the job (see BM p. 996), leading $Z_{j\ell}$ to be correlated with its leave out mean $\bar{Z}_{j\ell}$ and hence with $\bar{X}_{j\ell}$. The AGCV study includes only a small number of randomized resume characteristics that are not predictive of callback outcomes.

Table II: Tests for dependence

Nunley et al. data: resume characteristics			Arceo-Gomez & Campos-Vasquez: resume order					
	Main effect	Leave-out mean		Observations	χ^2 statistic	d.f.	P-value	Exact p-value
Variable	(1)	(2)	Callbacks	(3)	(4)	(5)	(6)	(7)
Black	-0.028	-0.019	*Panel A. Four-application sequences*					
	(0.010)	(0.027)	1	142	1.4	3	0.708	0.794
Female	0.010	0.009						
	(0.010)	(0.027)	2	99	10.0	5	0.075	0.155
High SES	-0.233	-0.674						
	(0.174)	(0.522)	3	64	3.2	3	0.367	0.513
GPA	-0.043	-0.153	*Panel B. Eight-application sequences*					
	(0.066)	(0.198)	1	56	7.8	7	0.347	0.504
Business major	0.008	0.010						
	(0.008)	(0.021)	2	37	23.6	27	0.651	0.697
Employment gap	0.011	0.034						
	(0.009)	(0.023)	3	36	58.4	55	0.352	0.397
Current unemp.: 3+	0.013	0.005						
	(0.012)	(0.032)	4	39	75.2	69	0.286	0.457
6+	-0.008	-0.038						
	(0.012)	(0.029)	5	16	40.7	55	0.924	1.000
12+	0.001	0.021						
	(0.012)	(0.032)	6	20	28.6	27	0.379	0.469
Past unemp.: 3+	0.029	0.065						
	(0.012)	(0.031)	7	6	8.4	7	0.300	0.539
6+	-0.011	-0.016	*Panel C. Joint tests*					
	(0.012)	(0.033)	Independence in all callback strata: χ^2 (247) = 242.7, p = 0.565					
12+	-0.004	0.019						
	(0.012)	(0.031)	No order effects: χ^2 (7) = 5.3, p = 0.622					
Predicted callback rate	0.476	-0.041						
	(0.248)	(0.626)	Regression of callback on order: coef. = -0.0021, s.e. = 0.0015, p = 0.147					
Joint p-value		0.452						
Sample size		9,220	Regression of callback on frac. females sent earlier: coef. = -0.003, s.e. = 0.013, p = 0.788					

Notes: This table reports results from tests of the assumption that applications at each job are independent Bernoulli trials with race-specific success probabilities. Columns (1) and (2) show tests based on resume characteristics using data from Nunley et al. (2015). Estimates come from regressions of a callback indicator on a resume characteristic and the mean of this characteristic across other resumes at the same job. The predicted callback rate is the fitted value from a regression of a callback indicator on all resume characteristics, leaving out the reference job. The joint p-value comes from a test of the hypothesis that coefficients on the leave-out mean are zero for all individual characteristics. Standard errors, clustered at the job level, appear in parentheses. Columns (3)-(7) show tests based on resume order in the Arceo-Gomez and Campos-Vasquez (2014) data. These results come from Wald tests of the hypothesis that all callback sequences leading to a particular total number of callbacks are equally likely. Column (3) shows the number of observed sequences in each callback stratum, column (4) shows the Pearson χ^2 goodness of fit statistic, column (5) shows the degrees of freedom for the test, column (6) shows the corresponding p-value, and column (7) shows an exact multinomial goodness of fit p-value obtained by summing probabilities of all sequence configurations that occur with probability less than or equal to the observed configuration under the null. Panel A constructs sequences separately for the first four and last four applications at each job, and panel B uses the full eight application sequence. The first two rows of Panel C show the results of a joint test of independence across all callback strata and a test that mean callback rates are equal across the eight resume order positions. The third row displays the slope coefficient from a linear regression of a callback indicator on order and a constant. The final row shows the slope coefficient from a regression of a callback indicator on the fraction of the four total female applications to the job that were sent prior to the reference application. This regression also includes a female indicator and a constant. Standard errors are clustered by job.

A second set of tests for independence exploits data on the specific order in which resumes were sent to jobs in the AGCV experiment (corresponding data were unavailable for BM and NPRS). With independent trials, all callback sequences leading to a particular total number of callbacks t should be equally likely, so each such sequence should constitute a share $\binom{L}{t}^{-1}$ of the sample calling t applications in total. Many plausible forms of dependence would manifest as violations of this condition. If employers stop calling after seeing enough high quality applicants, for example, we should expect to see sequences with runs of callbacks followed by non-callbacks. Likewise, if some employers detect the experiment after receiving several applications, we should see sequences with early callbacks overrepresented and fewer callbacks at later positions in the order.

Columns 3-7 of Table II provide tests of the independence assumption in each callback stratum t of the AGCV data. We form Pearson (1900) χ^2 test statistics equal to quadratic forms in the difference between observed and expected callback sequence frequencies, scaled by the covariance matrix of these differences under the null that each sequence in a stratum is equally likely. Panel A splits the sample of eight applications into two sequences of four at each job in order to increase

the expected frequency of each sequence, which may improve the power of the test against certain alternatives. Panel B displays results using the full eight-application sequence. These tests fail to reject the null hypothesis of independence in any callback stratum ($p \geq 0.07$) or across all strata jointly ($p = 0.57$). As shown in columns 7 and 8, the conclusions of this exercise are similar when we base inference on an asymptotic approximation to the distribution of the χ^2 statistic and when we use exact finite-sample p-values computed by summing probabilities of all sequence configurations that are less likely than the observed configuration under the null.

Tests based on the full set of callback sequences may have low power against some alternatives. Panel C of Table II therefore reports the results of three additional tests for specific plausible forms of dependence. A test that callback rates are equal across the eight resume order positions fails to reject ($p = 0.62$). Similarly, a linear regression of a callback indicator on resume order suggests that the callback rate declines slightly with order, but the slope coefficient is not statistically significant at conventional levels ($p = 0.15$). Finally, to test if earlier application influence an employer's perception of resume quality, the bottom row of panel C regresses a callback indicator on the share of the four total female applications to the job sent prior to the current application, controlling for whether the current application is female. This test again fails to reject the null of independence ($p = 0.79$). These results indicate that the independent Bernoulli trials model provides a good approximation to correspondence studies sending eight or fewer applications to each job. While we are not aware of any existing experiments sending more than eight applications to each job, a potentially interesting topic for future research is to study the nature of any dependence that arises as additional applications are sent to a given employer.

7 Moment and Posterior Estimates

To estimate the identified moments in each experiment we compute shape-constrained GMM (SCGMM) estimates that require the callback frequencies to be rationalizable by a proper discretized probability distribution defined on a 150×150 grid of support points. Imposing shape constraints serves two goals. First, we need the moment estimates to be rationalizable by some $G \in \mathscr{G}$ in order to subsequently use them as constraints when estimating bounds via our linear programming method. Second, when the constraints bind, the resulting estimates are typically closer to the truth and more precise (see Chetverikov et al., 2018 for a review). Details of the SCGMM estimation procedure, which involves solving a Quadratic Programming (QP) problem, appear in Appendix D.

Table III uses the shape constrained estimates to summarize key features of the distribution of callback probabilities in each experiment, and reports minimized SCGMM criterion functions (J-statistics) and p-values from bootstrap tests of the shape constraints based on the methods of Chernozhukov et al. (2015). The full set of unconstrained moment estimates appear in Appendix Tables A.I-A.III. Because the shape constraints may make the criterion non-differentiable, we rely

Table III: Non-parametric estimates of treatment effect variation in resume correspondence studies

Bertrand & Mullainathan			Nunley et al.			Arceo-Gomez & Campos-Vasquez		
p_b	p_w	$p_b - p_w$	p_b	p_w	$p_b - p_w$	p_m	p_f	$p_m - p_f$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
0.063	0.094	-0.031	0.153	0.177	-0.023	0.109	0.137	-0.028
(0.006)	(0.007)	(0.006)	(0.007)	(0.007)	(0.005)	(0.009)	(0.010)	(0.008)
0.152	0.199	0.082	0.290	0.308	0.102	0.229	0.257	0.178
(0.012)	(0.012)	(0.016)	(0.008)	(0.007)	(0.012)	(0.012)	(0.011)	(0.014)
0.927	1.00	-0.717	0.944	1.00	-0.336	0.738	1.00	-0.498
(0.051)	-	(0.119)	(0.017)	-	(0.066)	(0.039)	-	(0.058)
-	-	-	3.76	3.65	-4.45	4.04	3.74	-1.64
			(0.08)	(0.08)	(0.82)	(0.13)	(0.10)	(0.56)
-	-	-	-	-	-	8.59	5.91	13.6
						(1.13)	(0.71)	(3.5)
	0.0			23.1			2.7	
	1.00			0.190			0.891	

-constrained generalized method of moments (SCGMM) estimates of key features of the joint distribution of treatment and resume correspondence studies. Columns (1)-(3) show estimates for black and white callback rates in Bertrand and Mullainathan ay estimates for black and white callback rates in Nunley et al. (2015), and columns (7)-(9) show estimates for male and female z and Campos-Vasquez (2014). Standard errors are computed using the numerical bootstrap procedure described by Hong and are minimized SCGMM criterion functions. *P* -values come from bootstrap tests of the hypothesis that the model restrictions are

on the "numerical bootstrap" procedure of Hong and Li (forthcoming) to construct pointwise valid estimates of standard errors.[10]

Table IV reports LP estimates of the lower bound probability that a given employer is discriminating. In computing both the analytic bounds of Lemma 2 and the sharp bounds of (5), we replace the unknown callback probabilities $\bar{f}$ with estimates $\hat{\bar{f}} = B\hat{\mu}$, where $\hat{\mu}$ is the relevant vector of shape-constrained moment estimates produced by our SCGMM procedure. Because the LP algorithm used to solve (5) scales efficiently to large problems, we use a finer discretization with 36 times as many points as the grid used in our earlier SCGMM step.[11]

Bertrand and Mullainathan (2004)

The first rows of columns 1 and 2 of Table III show the mean callback probabilities of white and black applications across jobs. The J-statistic of zero reported in column 2 of Table III indicates that the shape constraints do not bind in the BM data, i.e. that the sample frequencies can be rationalized to numerical precision by a discretized probability distribution. Because the shape constraints do not bind and the BM application design is balanced, the mean callback probabilities match the callback rates reported in Table I perfectly. More interesting are the second moments: there is substantial over-dispersion in callback probabilities, with standard deviations across jobs for each race-specific probability more than double the mean probability. As expected, there is also a strong positive correlation between white and black callback rates, reflecting that some employers simply call back more applications of all types.

Column 3 of Table III reveals substantial heterogeneity in the difference in race specific callback rates $p_{jb}-p_{jw}$ across jobs, with a standard deviation more than twice as large as the mean. The third row shows a strong negative correlation between the discriminatory gap in callback rates $p_{jb} - p_{jw}$ and the white callback probability p_{jw}, suggesting that discrimination tends to be stronger when jobs have higher chances of calling back more white workers. This reflects, in part, a mechanical boundary effect, as an employer with very low callback rates has little opportunity to discriminate. Since the white callback rate in this study is only around 10%, boundary effects are likely to be a quantitatively important phenomenon.

Column 1 of Table IV reports lower bounds on the share of jobs engaged in discrimination by the number of total callbacks in the BM experiment. The analytic bounds in Lemma 2 (presented in brackets) imply that at least 38% of the jobs that call back 2 applications are engaged in discrimination, while at least 44% of jobs that call back 3 applications are discriminating. The sharp LP bounds are somewhat tighter than their analytical counterparts, revealing that at least 44% of the jobs calling back two applicants are discriminating. Among jobs that call back three

[10] Because the asymptotic distribution of the shape constrained estimator will tend to be non-normal (Fang and Santos, 2018), standard errors provide only a heuristic guide to the uncertainty associated with each moment estimate.

[11] Appendix Table A.IV assesses the sensitivity of our estimates to alternative discretization schemes. The results show that the moment estimates are not sensitive to the number of grid points used in the SCGMM step (as evidenced by the goodness of fit statistic) and that the bounds stabilize with a sufficiently large number of grid points in the linear programming step.

Table IV: Lower bounds on probabilities of discrimination

Bertrand & Mullainathan			Nunley et al.			Arceo-Gomez & Campos-Vasquez		
$\neq p_b$)	Pr($p_w < p_b$)	Pr($p_b < p_w$)	Pr($p_w \neq p_b$)	Pr($p_w < p_b$)	Pr($p_b < p_w$)	Pr($p_f \neq p_m$)	Pr($p_f < p_m$)	Pr($p_m < p_f$)
)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
0	0.000	0.130	0.358	0.154	0.173	0.207	0.064	0.142
8	0.000	0.038	0.152	0.093	0.048	0.065	0.023	0.042
4	0.000	0.424	0.672	0.185	0.433	0.721	0.307	0.414
3}			{0.176}			{0.071}		
2	0.000	0.442	0.691	0.016	0.675	0.708	0.226	0.481
9}			{0.282}			{0.524}		
8	0.000	0.508	0.821	0.067	0.736	0.584	0.050	0.533
0}			{0.126}			{0.502}		
2	0.000	0.212	0.421	0.257	0.128	0.518	0.053	0.465
						{0.492}		
						0.320	0.153	0.167
						{0.286}		
						0.372	0.176	0.197
						{0.308}		
						0.453	0.122	0.331
						{0.170}		
						0.069	0.008	0.062
3	0.0	29.3	62.6	23.5	62.6	427.8	27.1	421.0
0	1.00	0.000	0.000	0.120	0.000	0.000	0.018	0.000

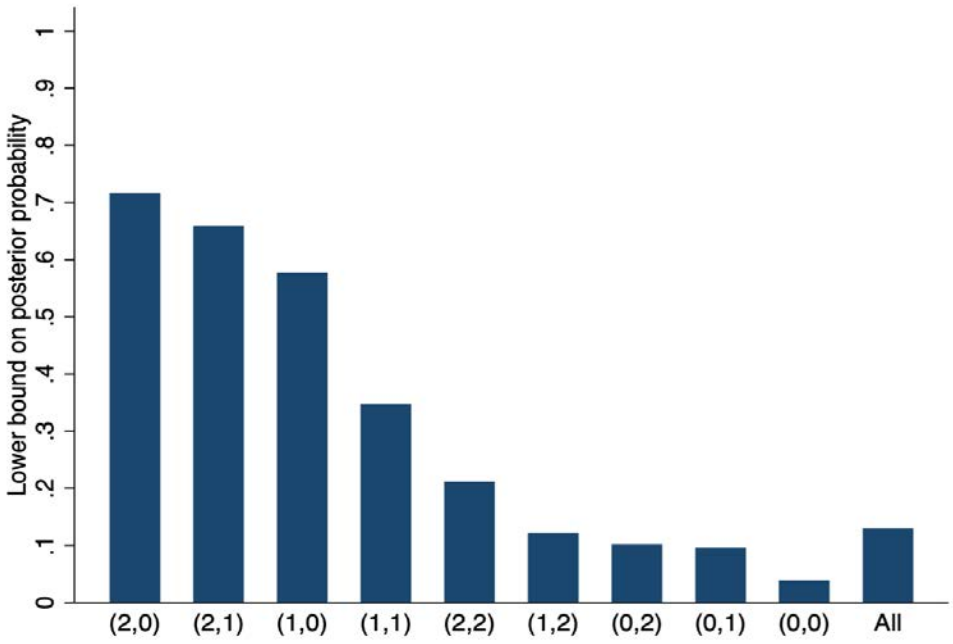
s lower bounds on the probability that jobs discriminate based upon race or sex. Bounds are computed via linear ossible, corresponding analytic bounds based on the formula in Lemma 2 appear in brackets. The first row shows bounds in s, and the remaining rows display bounds conditional on the total number of callbacks. Columns (1)-(3) show results for the Bertrand and Mullainathan (2004) data, while columns (4)-(6) show results for racial discrimination in the Nunley et al. 1) and (4) display lower bounds on the fraction of jobs with equal callback rates for white and black applicants, columns (2) unds on the fraction discriminating against white applicants, and columns (3) and (6) report lower bounds on the fraction lack applicants. Results for the Nunley et al. (2015) data that condition on the number of callbacks refer to jobs receiving two lications. Columns (7)-(9) show results for sex discrimination in the Arceo-Gomez and Campos-Vasquez (2014) data. wer bound on the fraction of jobs with equal callbacks for men and women, column (8) shows a lower bound on the fraction omen, and column (9) reports a lower bound on the fraction discriminating against men. J-statistics and p-values come he hypothesis that the lower bound equals zero for all jobs.

applications, at least half are discriminating on the basis of race. Hence, in this callback stratum, our estimates suggest jobs should not logically be presumed "innocent" of discrimination.

The LP approach also generates informative bounds in callback strata for which analytical bounds are not available. Overall, at least 13% of jobs discriminate on the basis of race. Notably, at least 4% of jobs that call back no applications are engaged in discrimination, while at least 21% of jobs that call back all four applications discriminate on the basis of race. Since neither of these strata exhibit any difference in black-white callback rates, all of the relevant information on discrimination in these strata comes from the total number of callbacks blended with the indirect evidence from the population distribution G.

Column 2 of Table IV reports LP-based lower bounds on the proportion of jobs with white callback probabilities less than their black callback probabilities. We find a lower bound of exactly zero in each callback stratum, indicating that the callback probabilities can be rationalized without any employers engaging in "reverse discrimination" against whites. Column 3 reports lower bounds on the proportion of jobs with white callback probabilities greater than their black callback probabilities. These lower bound estimates coincide exactly with those reported in column 1. Accordingly, we easily reject the null hypothesis of no discrimination against blacks.

Figure I: Lower bounds on posterior probabilities of discrimination, BM data



Notes: This figure displays lower bounds on the probability that jobs in the Bertrand and Mullainathan (2004) data discriminate based upon race given their callback configurations. Each bar reports a lower bound on the posterior probability of discrimination conditional on (C_w, C_b), where C_w is the number of white callbacks and C_b is the number of black callbacks.

Figure I converts the lower bound estimates in column 3 of Table IV to lower bound posterior probabilities of discrimination. Overall, at least 13% of jobs engage in discrimination. However, at least 72% of jobs that call back two white and no black applications are discriminating, while a job

that calls back one white and no black applications has at least a 58% chance of being engaged in discrimination.

Nunley et al. (2015)

Moment estimates from the NPRS study are reported in columns 4-6 of Table III. Recall that NPRS employed five distinct application designs with $(L_{jw}, L_{jb}) \in \{(4,0),(3,1),(2,2),(1,3),(0,4)\}$. Appendix Table A.II reports design-specific method of moments estimates of all identified moments for the three designs with the largest sample sizes.[12] As expected, the design-specific estimates are generally close to one another and we cannot reject that they are identical. To pool the designs efficiently, we again use an SCGMM estimator that requires the moments be rationalizable by a proper probability distribution $G \in \mathscr{G}$. The minimized SCGMM criterion function provides a measure of the goodness of fit of our model. Applying the bootstrap method of Chernozhukov et al. (2015) yields a p-value of 0.19 for the null hypothesis that the results for all experimental designs are jointly rationalized by a common distribution G.

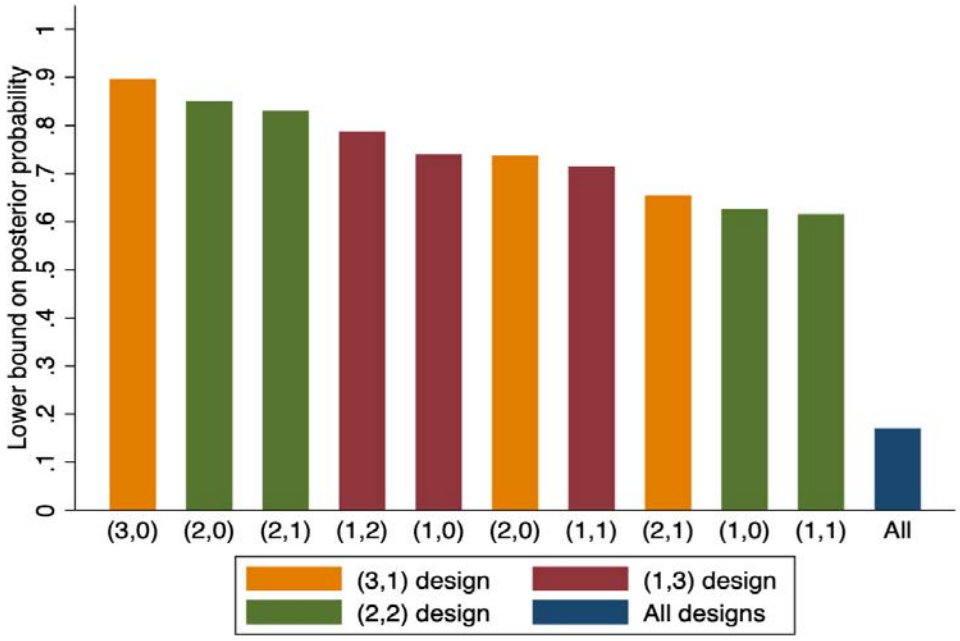
Consistent with our findings for the BM data, columns 4-6 of Table III reveal substantial heterogeneity in race-specific callback rates in the NPRS experiment, with standard deviations roughly twice their mean. The imbalanced designs used by NPRS allow us to identify higher moments than the earlier BM study even though the two studies sent the same number of applications per job. While race-specific callback rates are right skewed, racial gaps in callback probabilities $p_{jb} - p_{jw}$ are left-skewed, indicating a long tail of heavy discriminators.

Columns 4-6 of Table IV report estimated lower bounds on the probability of discrimination from the NPRS study for the full population of jobs as well as bounds conditional on total callbacks in a balanced design with $L_{jw} = L_{jb} = 2$. In column 1, our analytic bound formula suggests at least 28% of the jobs calling back two applicants in this design are discriminating – slightly lower than the corresponding estimate in BM. Applying the LP approach tightens the analytic bounds dramatically and provides additional bounds on the prevalence of discrimination among jobs that make no callbacks or that call every application. We estimate that at least 36% of all jobs have different white and black callback probabilities, with that share rising to 69% among employers who call back two applicants in a balanced $(2,2)$ design.

Some of this discrimination is estimated to be against whites. Column 5 shows that the shape constrained callback probabilities $\hat{\bar{f}}$ imply that at least 15% of employers have white callback probabilities less than their black probabilities. These moments are estimated with error, however, and a bootstrap test of the null hypothesis that all employers have white callback probabilities weakly exceeding their black callback probabilities yields a p-value of 0.12. If we attribute the evidence of reverse discrimination to sampling error, we can take the estimates in column 6 as the relevant lower bounds on discrimination. These results imply that at least 17% of jobs discriminate against black applicants. We decisively reject the null hypothesis that this lower bound is zero.

[12] The remaining designs were omitted from this analysis due to small sample sizes. Only 22 jobs were in the $(L_{jw} = 0, L_{jb} = 4)$ design while 43 jobs fell in the $(L_{jw} = 4, L_{jb} = 0)$ design.

Figure II: Lower bounds on posterior probabilities of discrimination, NPRS data



Notes: This figure displays lower bounds on the probability that jobs in the Nunley et al. (2015) data discriminate against black applicants for the 10 callback configurations with highest posterior bounds. Each bar reports a lower bound on the posterior probability that $p_w > p_b$ conditional on (C_w, C_b), where C_w is the number of white callbacks and C_b is the number of black callbacks. Orange bars correspond to an experimental design with 3 white and 1 black application, green bars correspond to a design with 2 white and 2 black applications, and red bars correspond to a design with 1 white and 3 black applications. The blue bar reports the lower bound on the prior probability of discrimination.

Figure II converts these lower bound priors into posterior estimates of the share of employers with selected callback configurations engaged in discrimination against black applicants. We estimate that at least 85% of the employers calling back two white and no black applicants in a balanced $(2, 2)$ design are discriminating against blacks. Interestingly, calling back three whites and no blacks in a $(3, 1)$ design is estimated to be even more suspicious, with at least 90% of the employers generating this callback evidence engaged in discrimination against black applicants.

Arceo-Gomez and Campos-Vasquez (2014)

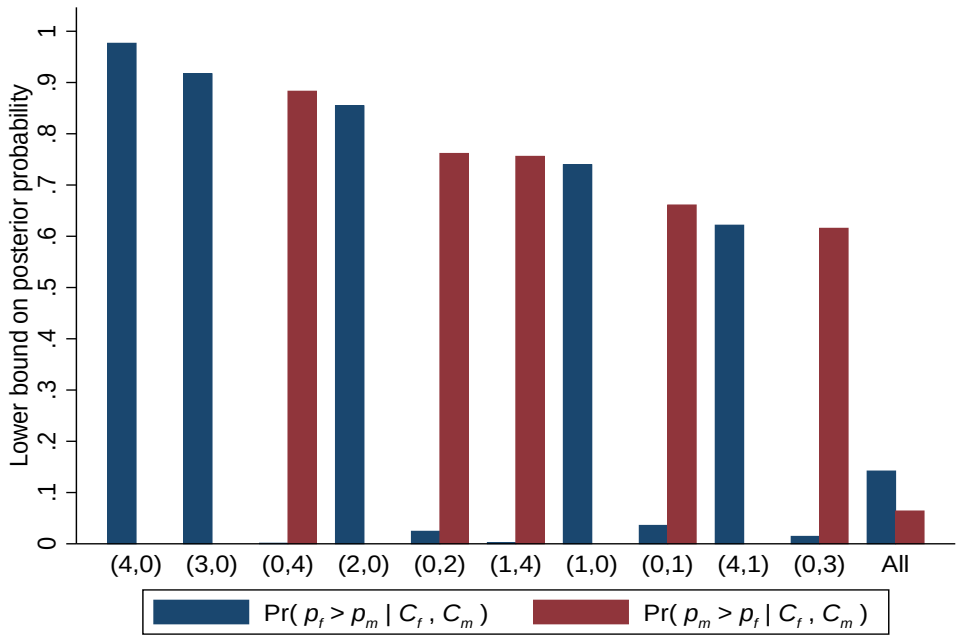
The full set of moment estimates for the AGCV experiment (reported in Appendix Table A.III) reveal that the shape constraints bind strongly in this case, presumably because the design of the AGCV experiment involves many small cells. Despite substantial movement in some moment estimates, the bootstrap p-value from a test of the null hypothesis that the callback frequencies are generated by the model is 0.89, indicating that the raw callback frequencies are rationalizable by a well-behaved underlying joint distribution of callback probabilities.

Columns 7-9 of Table III report key moment estimates from the AGCV data. The behavior of the first two moments is similar to that reported in the prior two experiments, with gender-specific

standard deviations roughly twice their mean callback probabilities. However, the greater number of applications used in this design helps enormously with the precision of higher moment estimates. We find strong evidence of left-skew in the distribution of gender gaps in callback probabilities as well as evidence of excess kurtosis in the distribution of gaps. While many jobs discriminate little, there is a thick tail of heavy discriminators.

Columns 7-9 of Table IV report estimated lower bounds on the probability of discrimination in the AGCV experiment. Focusing on the sharp bounds reported in column 7, we find that at least 21% of jobs are engaged in discrimination. Remarkably, this share rises to 72% among jobs calling back one applicant and 71% among jobs calling two. These shares are much higher than the corresponding analytic bounds, showing that cross-stratum restrictions in a design with eight applications are useful for tightening bounds in strata with few callbacks. Evidently, jobs that call back few applicants in the AGCV experiment are very likely to engage in discrimination.

Figure III: Lower bounds on posterior probabilities of discrimination, AGCV data



Notes: This figure displays lower bounds on the probability that jobs in the Arceo-Gomez and Campos-Vasquez (2014) data discriminate based upon sex for the 10 callback configurations with highest posterior bounds. Each bar reports a lower bound on the posterior probability of discrimination conditional on (C_f, C_m), where C_f is the number of female callbacks and C_m is the number of male callbacks. Blue bars report lower bounds on the probability of discriminating against men, and red bars report lower bounds on the probability of discriminating against women.

Some of this discrimination appears to be "reverse" discrimination against women. Column 8 shows that at least 6% of jobs discriminate against women and a bootstrap test of the null hypothesis that this bound equals zero is decisively rejected ($p < 0.02$). An employer that calls back a single application has at least a 31% chance of discriminating against women. Column 9 shows that at least 14% of jobs discriminate against men, and the bootstrap p-value indicates

From callbacks to labor markets - Kline & Walters, Reasonable Doubt (2021) 599

this bound is also statistically distinguishable from zero. The mean difference in callback rates in the ACGV experiment therefore masks gender discrimination operating in both directions. An employer that calls back a single application has at least a 41% chance of discriminating against men.

Figure III plots lower bound posterior probabilities of discrimination against men and women, respectively, for selected callback configurations. At least 97% of the jobs that call back four women and no men are estimated to discriminate against men, and at least 88% of jobs that call back four men and no women are estimated to discriminate against women. But even an employer that calls back a single woman and no men has at least a 74% chance of discriminating against men. Likewise, at least 66% of jobs that call back a single man and no women are estimated to discriminate against women. That we obtain such informative posteriors in settings with a single callback demonstrates the tremendous value of indirect evidence in this setting.

8 Experimental Design and Detection Error Tradeoffs

The above analysis demonstrated that it is possible to achieve high posterior certainty that individual jobs are engaged in discrimination when callback rates at those jobs differ dramatically across protected groups. Can such evidence be used to reliably detect a non-trivial share of discriminating jobs? We address this question by studying the tradeoff between Type I and II errors that arises under a simple two-type mixture specification calibrated to match callback rates in the NPRS data given race and other resume characteristics. We then consider how the resulting detection error tradeoffs change as the experimental design is altered to send more applications to each job.

A Mixed Logit Model

We work with a mixed logit model for callbacks of the form:

$$\Pr\left(Y_{j\ell}=1|R_{j\ell},X_{j\ell},\alpha_j,\beta_j\right)=\Lambda\left(\alpha_j-\beta_j 1\left\{R_{j\ell}=b\right\}+X_{j\ell}'\psi\right),$$

where $\Lambda(\cdot)=\frac{\exp(\cdot)}{1+\exp(\cdot)}$ is the standard logistic CDF, $X_{j\ell}$ is a vector of de-meaned application covariates, and (α_j,β_j) are random coefficients governing the odds of a white callback and discrimination against blacks, respectively. To allow for heterogeneity in white callback rates we assume that $\alpha_j \overset{iid}{\sim} N\left(\alpha_0,\sigma_\alpha^2\right)$. Discrimination is modeled as a two-type (conditional) mixture:

$$\beta_j|\alpha_j=\begin{cases}\beta_0 & \text{w/ prob. } \Lambda\left(\tau_0+\tau_\alpha\alpha_j\right),\\ 0 & \text{w/ prob. } 1-\Lambda\left(\tau_0+\tau_\alpha\alpha_j\right).\end{cases}$$

This specification allows for some share of jobs to not discriminate at all, while the remaining jobs depress the odds of calling back blacks relative to whites by roughly $\beta_0\%$. When $\tau_\alpha\neq 0$, the probability of discrimination depends on α_j, which governs the white callback rate. Note that

random assignment of the covariates $X_{j\ell}$ implies they are independent of (α_j, β_j) and therefore excludable from the type probability equation.

Model Estimates

Table V shows the results of fitting the above model to the NPRS experiment by simulated maximum likelihood. Column 1 provides a standard "random effects" logit model with heterogeneity confined to the intercept as in Farber et al. (2016). We find substantial variability across jobs in the overall odds of a callback: a 0.1 standard deviation increase in the intercept α_j is estimated to raise the odds of a callback by 47%. We also find clear evidence of market-wide discrimination: black applications have roughly 46% lower odds of being called back than their white counterparts.

Table V: Mixed logit parameter estimates, NPRS data

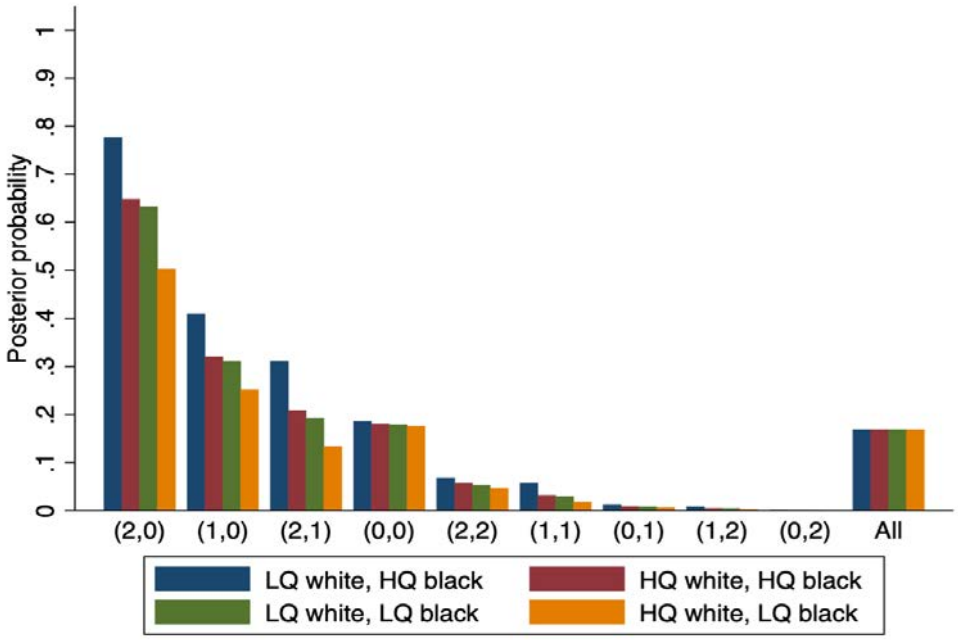
		Types	
	Constant	No selection	Selection
	(1)	(2)	(3)
Distribution of logit(p_w): α_0	-4.71	-4.93	-4.93
	(0.22)	(0.24)	(0.28)
σ_α	4.74	4.99	4.98
	(0.22)	(0.25)	(0.29)
Discrimination intensity: β_0	0.456	4.05	4.05
	(0.108)	(1.56)	(1.58)
Discrimination logit: τ_0	-	-1.59	-1.56
		(0.42)	(1.10)
τ_α	-	-	-0.005
			(0.180)
Fraction with $p_w \neq p_b$:	1.00	0.168	0.170
Log-likelihood	-2,792.1	-2,788.2	-2,788.2
Parameters	15	16	17
Sample size	2,305	2,305	2,305

Notes: This table reports simulated maximum likelihood estimates of mixed logit models for callback probabilities in the Nunley et al. (2015) data. Columns (2)-(3) allow for two discrete types of firms, one of which does not discriminate based upon race. All models include resume covariates. Covariates are de-meaned in the estimation sample. Robust standard errors in parentheses.

Column 2 allows the race effect β_j to vary across employers, which yields a significant improvement in model fit. The types specification finds that only about 17% of jobs discriminate against blacks – very near the non-parametric lower bound estimate produced earlier by our LP routine (see column 6 of Table IV). The degree of discrimination among such jobs is estimated to be severe: the odds of receiving a callback are roughly $\exp(4) - 1 \approx 53$ times higher for white applications than for blacks. Column 3 allows the probability of discrimination to vary with the white callback rate,

which yields a negligible improvement in model fit. Surprisingly, α_j and β_j are found to be nearly independent, which implies that the negative correlation between $p_{jb}-p_{jw}$ and p_{jw} reported in Table III is attributable to boundary effects. Again, this model finds roughly 17% of jobs discriminate against blacks. Because we cannot reject the null hypothesis that $\tau_\alpha = 0$, we work with the more parsimonious model in column 2 in the exercises that follow. Appendix Figure A.I provides a goodness of fit diagnostic for this model, plotting the empirical callback rates in each black / white callback by application design cell against the logit model's predicted callback probability in that cell. The empirical frequencies track the model predictions closely and a naive Pearson χ^2 test fails to reject the null hypothesis that the model rationalizes the cell frequencies up to sampling error.

Figure IV: Mixed logit estimates of posterior discrimination probabilities, NPRS data



Notes: This figure displays mixed logit estimates of the posterior probability that jobs in the Nunley et al. (2015) data discriminate against black workers conditional on (C_w, C_b), where C_w is the number of white callbacks and C_b is the number of black callbacks. Blue bars show posteriors for a design sending two low quality (LQ) white applications and two high quality (HQ) black applications, where low and high quality are defined based on a logit covariate index 1 standard deviation below or above the mean. Red bars show posteriors for a design sending two HQ white and two HQ black applications. Green bars show posteriors for a design sending two LQ white and two LQ black applications. Orange bars show posteriors for a design sending two HQ white and two LQ black applications.

Posteriors

Figure IV reports the distribution of posterior probabilities $\Pr(D_j = 1|\{Y_{j\ell}, R_{j\ell}, X'_{j\ell}\psi\}_{\ell=1}^{L})$ implied by the parameter estimates reported in column 2 of Table V. To summarize the influence of the covariates, we evaluate the posteriors at two points within each race group, corresponding to the estimated index $X'_{j\ell}\hat{\psi}$ being a standard deviation above or below its empirical mean, which we refer to as "high" and "low" quality applications. By construction, the mean posterior coincides with

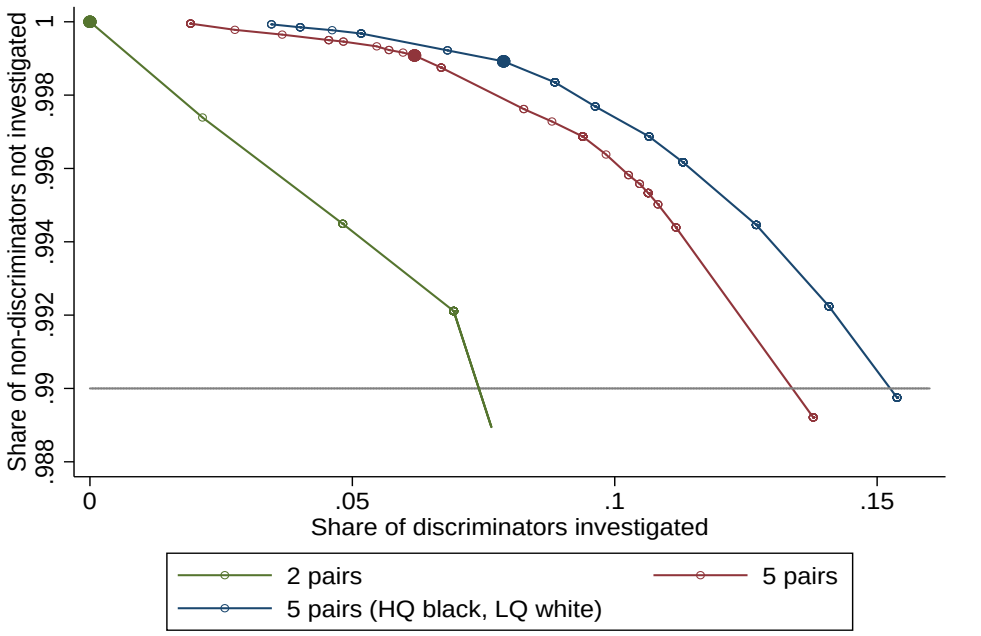
the estimated share of jobs that discriminate. The types model finds that only 17% of jobs are discriminating, yielding a strong prior that the typical job is not violating employment law. Yet calling back only white applicants still justifies a substantial degree of suspicion: 62% of the jobs that call back two whites and no blacks are discriminating.

Imbalances in the covariate mix of applicants can substantially intensify this suspicion. For example, 79% of the jobs that call back two low quality white applications and neither of two high quality black applications are discriminating. Evidently, even in models with a strong presumption of innocence, four applications can provide enough information to cast substantial doubt on whether individual employers are in compliance with employment law. However, it is only under the most extreme callback configurations that we can detect discriminators with reasonable certainty.

Detection Error Tradeoffs

Consider now a hypothetical regulator who forms posteriors taking as prior knowledge the two-type estimates reported in Table V. One may think of the regulator as first learning the parameters of the two-type model from a large experiment and then sending applications to additional vacancies drawn from the same population from which the original study sampled.

Figure V: Detection/error tradeoffs, NPRS data



Notes: This figure displays detection/error tradeoff curves based on models fit to the Nunley et al. (2015) data. Estimates come from decision rules applied to experiments generated from the logit model in column (2) of Table V. The horizontal axis measures the share of discriminating jobs investigated by each decision rule, while the vertical axis measures the share of non-discriminating jobs not investigated. The curves are generated by varying the posterior threshold at which jobs are investigated. The green curve corresponds to an experiment that sends two white and two black applications to each job, and the red curve corresponds to sending five applications of each race. These two curves randomly assign a 2-valued covariate index of resume quality (high or low), defined as +/-1 the empirical standard deviation of this index. The blue curve shows results from sending five low-quality white and five high-quality black applications. Bold points correspond to 80% posterior thresholds.

Figure V displays a rescaling of the Type I and II error rates that arise from investigating all jobs exceeding various posterior thresholds. The horizontal axis gives the share of jobs engaged in discrimination that are investigated. The vertical axis plots the share of non-discriminators that are not investigated, which can equivalently be viewed as one minus the False Discovery Rate. Each point gives the values of these shares corresponding to a particular posterior decision threshold. The bold point corresponds to a posterior threshold of 80%.

In the canonical design with only two white and two black applications per job, the 80% posterior threshold yields almost no false accusations. This control over Type I errors comes at the cost of a high Type II error rate – few accusations of any sort are made, leading to a negligible share of discriminators detected. Note that conducting a classical hypothesis test (e.g., Fisher's exact test) at the 1% level is equivalent to controlling the share of non-discriminators that are not investigated, which is depicted by the horizontal line at 0.99. This rule would yield more investigations but most of these would be erroneous: the equivalent posterior threshold in the 2 pair design is only 33%.

Expanding the design to 5 pairs of applications yields a substantial outward shift in the detection error tradeoff curve. Using a posterior threshold of 80% keeps the share of employers erroneously investigated for discrimination below 0.2% while allowing detection of roughly 7.5% of jobs that discriminate. Lowering the posterior threshold further boosts the detection rate above 10% while modestly increasing the Type I error rate. This shows that ten applications is enough to accurately detect a non-trivial share of discriminators.

The third line shows the results of an experiment where each job is sent 5 high quality black applications and 5 low quality white applications.[13] Modifying the experimental design in this way yields additional improvements in Type I and II error rates. Using an 80% posterior threshold, the share of non-discriminators investigated remains below 0.2%, while the share of discriminators investigated rises to almost 10%.

9 Indirect Evidence and Policy

The findings of the previous section indicate that correspondence experiments with as few as 10 applications can reliably detect a substantial share of discriminating employers when the population distribution of callback probabilities is known. We now consider how a Bayesian regulator might go about deciding which jobs to investigate and then assess how partial identification of the callback rate distribution affects this decision rule.[14]

[13] Of course, the results of such an experiment would be difficult to interpret without an earlier experiment revealing the model parameters, as one would not be able to parse the effects of race from those of quality.

[14] Our analysis is based loosely on the functioning of the EEOC, which has the authority to conduct systematic investigations into the discriminatory behavior of particular organizations. Because investigations are costly, the EEOC uses a priority system based on human judgement to decide which complaints to investigate (US Equal Employment Opportunity Commission, 2016). The results below illustrate how direct callback evidence from a correspondence experiment can be blended with indirect evidence to assist or replace informal human judgements.

The Regulator's Problem

Suppose the regulator must decide which jobs to investigate based on a vector of direct evidence $\mathcal{E}_j = \{Y_{j\ell}, R_{j\ell}, X_{j\ell}\}_{\ell=1}^{L}$ revealed by a correspondence study. The regulator uses a deterministic decision rule $\delta(\mathcal{E}_j)$ that maps this evidence vector to a binary inquiry decision.[15] Each job has a pair of race specific callback probabilities $\{p_{jw}(x), p_{jb}(x)\}$ that may vary with applicant quality x. We use $H(\cdot)$ to denote the *iid* randomization distribution of $X_{j\ell}$. Consistent with our earlier analysis of the NPRS experiment, we rule out the possibility of discrimination against whites by assuming $\Pr(p_{jw}(x) \geq p_{jb}(x)) = 1$ for all quality levels $x \in \mathcal{X}$.

The regulator's loss from applying decision rule $\delta(\mathcal{E}_j) = \delta_j \in \{0, 1\}$ to job j is modeled as:

$$\mathcal{L}_j(\delta_j) = \delta_j \left(\kappa - \Lambda \left(\int \left[\Lambda^{-1}(p_{jw}(x)) - \Lambda^{-1}(p_{jb}(x)) \right] dH(x) \right) \right). \tag{8}$$

One can think of the parameter $\kappa \in (1/2, 1]$ as capturing the cost of conducting an investigation. The term $\Lambda\left(\int \left[\Lambda^{-1}(p_{jw}(x)) - \Lambda^{-1}(p_{jb}(x))\right] dH(x)\right) \in [1/2, 1]$ gives the benefit to the investigation, which is increasing in the average log callback odds advantage of whites over blacks at job j across quality levels. To ensure this benefit remains bounded, the racial difference in log odds is then mapped back to the unit interval by the logistic CDF to produce the payoff to an investigation. Note that under the logit model of the previous section this payoff reduces to $\Lambda(\beta_j)$. The regulator would like to investigate whenever this payoff exceeds the investigation cost, in which case $\mathcal{L}_j(1)$ is negative.

Because the $\{p_{jw}(x), p_{jb}(x)\}_{x \in \mathcal{X}}$ are not known, the regulator minimizes expected loss (i.e. risk). When the regulator knows the joint distribution of callback probabilities in the population, the risk function can be written:

$$\mathcal{R}_j(G, \delta(\cdot)) = \mathbb{E}\left[\mathcal{L}_j(\delta_j)\right] = \mathbb{E}\left[\delta_j(\mathcal{E}_j) \left(\kappa - \Lambda \left(\int \left[\Lambda^{-1}(p_{jw}(x)) - \Lambda^{-1}(p_{jb}(x)) \right] dH(x) \right) \right)\right],$$

where $G : [0,1]^{2|\mathcal{X}|} \to [0,1]$ is the joint distribution of quality-specific callback rates. Because $\{\mathcal{E}_j, p_{jw}(x), p_{jb}(x)\}_{j=1}^{J}$ are *iid* across jobs, we can drop the j subscripts and refer to the risk function as $\mathcal{R}(G, \delta(\cdot))$. Choosing $\delta(\mathcal{E}_j)$ to minimize the risk function pointwise yields the following Lemma, which characterizes the regulator's optimal decision rule in the case where G is known.

Lemma 3 (Optimal Decision Rule)**.**

$$\delta(\mathcal{E}_j) = 1\left\{ \mathbb{E}\left[\Lambda\left(\int \left[\Lambda^{-1}(p_{jw}(x)) - \Lambda^{-1}(p_{jb}(x)) \right] dH(x) \right) | \mathcal{E}_j \right] > \kappa \right\} \text{ minimizes } \mathcal{R}(G, \delta).$$

One can think of Lemma 3 as offering an economically motivated standard of *reasonable doubt*: when the posterior expected benefit of an investigation exceeds the investigation cost κ, it is rational to conduct an investigation. Note that in the two-type logit model the difference in log odds at

[15] We confine ourselves to deterministic rules because randomized decision rules violate commonly held horizontal equity principles.

job j equals $\beta_0 D_j$ for all quality levels, so the optimal decision rule amounts to investigating when the posterior probability of discrimination exceeds a cost-based threshold. In the logit model, for example, the decision rule can be written $\delta(\mathcal{E}_j) = 1\{\mathcal{P}(\mathcal{E}_j, G_{logit}) > (\kappa - 1/2)/(\Lambda(\beta_0) - 1/2)\}$.

Ambiguity

When G is only known to lie in some identified set Θ of distributions, many possible decision rules are consistent with rationality. Among those rules, an important benchmark is the minimax decision rule (Wald, 1945; Savage, 1951; Manski, 2000), which minimizes the maximum risk that may arise from the regulator's decisions. We can define the maximum risk function and the associated minimax decision rule respectively as:

$$\mathcal{R}^m(\Theta, \delta) = \sup_{G \in \Theta} \mathcal{R}(G, \delta) \quad \text{and} \quad \delta^{mm} = \arg\inf_{\delta \in \mathscr{D}} \mathcal{R}^m(\Theta, \delta), \tag{9}$$

where $\mathscr{D}$ is the set of deterministic decision rules. Unlike in the case where G is known, a regulator that only knows $G \in \Theta$ cannot consult a single posterior expectation to make the decision of whether to investigate. Rather, the maximum risk of each decision rule must be computed to obtain the minimax decision rule.

Relying on a discretized function space for G simplifies computation of the maximum risk function $\mathcal{R}^m$ consistent with a set of experimental callback probabilities. As explained in Appendix E, when Θ consists of a family of discrete distributions, $\mathcal{R}^m(\Theta, \delta)$ can be computed numerically as the solution to a linear programming problem. The minimax decision rule $\delta^{mm}(\cdot)$ is found by computing $\mathcal{R}^m(\Theta, \delta)$ for each candidate rule $\delta \in \mathscr{D}$ and choosing the rule that yields lowest maximal risk.

Bayes vs. Minimax Decisions

We now compare the decisions made by a Bayesian regulator with a minimax regulator in the hypothetical 5-pair generalization of the NPRS experiment considered in Section 8. As in Figure IV, we assume applications take on only two quality levels (high or low) with equal probability.

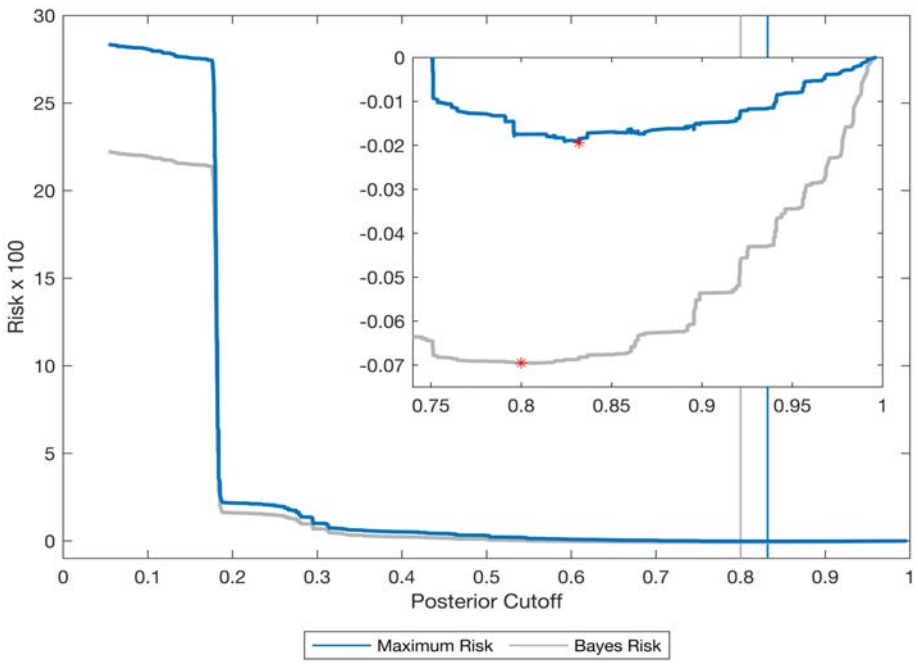
We consider a restricted family $\mathscr{D}^\dagger \subset \mathscr{D}$ of decision rules of the form $\delta(\mathcal{E}_j) = 1\{\mathcal{P}(\mathcal{E}_j, G_{logit}) \geq q\}$, where $q \in (0, 1)$ is a posterior cutoff and G_{logit} is the logit model reported in column 2 of Table V. Computing the maximal risk for this family of decision rules can be thought of as a way of "second guessing" the risk associated with each logit posterior threshold without debating the logit model's ordering of the underlying evidence configurations.[16] In computing $\mathcal{R}^m(\Theta, \delta)$, we use the logit model predictions of callback probabilities within each of the two quality bins as constraints (see the Appendix for details) and calibrate κ so that, under the logit DGP, an 80% posterior threshold minimizes risk.

[16] With L_w white and L_b black applications there are $2^{(1+L_w)(1+L_b)}$ logically possible decision rules which, in practice, prohibits brute force enumeration when $L_w + L_b > 4$. Restricting attention to logit posterior threshold rules allows us to circumvent this obstacle. In cases where multiple evidence configurations yield the same logit posterior, we consider separate rules that investigate each of these configurations individually.

From callbacks to labor markets - Kline & Walters, Reasonable Doubt (2021) 606

Figure VI plots logit (i.e., Bayes) risk and $\mathcal{R}^m(\Theta, \delta)$ against the nominal logit posterior threshold q. As q approaches one, both the maximal and Bayes risks approach zero, as no jobs are investigated in the limit. Conversely as the posterior threshold approaches zero – at which point all jobs are investigated – the maximum risk diverges from the Bayes risk because the least favorable G is one where nearly all jobs are engaged in trivial levels of discrimination that fail to justify the investigation cost. Recall from Table IV, however, that some jobs in the NPRS experiment must be discriminating, which limits the magnitude of this divergence.

Figure VI: Bayes and minimax risk, NPRS data



Notes: This figure displays risk functions generated by a hypothetical experiment sending five white and five black applications to jobs in the population studied by Nunley et al. (2015). Resumes are randomly assigned to high or low quality with equal probability, where quality is defined as +/-1 the empirical standard deviation of the logit covariate index. The horizontal axis plots the posterior threshold at which jobs are accused of discrimination. The grey curve displays risk based on the logit data generating process in column (2) of Table V. The cost of an investigation is calibrated so that a Bayesian regulator sets the posterior cutoff at 0.8. The blue curve plots maximum risk calculated by choosing the joint distribution of callback probabilities to maximize risk for each decision rule subject to the moments identified by the Nunley et al. (2015) experiment, restricted to decision rules that order jobs by the logit posterior threshold. Vertical lines indicate risk-minimizing thresholds. The window displays logit and minimax risk at threesholds of 0.75 and above, with risk-minimizing points indicated in red.

While the Bayes risk function is minimized by the decision rule with a posterior threshold of 80%, $\mathcal{R}^m(\Theta, \delta)$ is minimized by a rule with a logit-based threshold of 83%. This higher threshold implies a minimax regulator would investigate fewer jobs than a Bayesian regulator with the same preferences who believes G to be logit. The change in behavior is minimal, however, suggesting

From callbacks to labor markets - Kline & Walters, Reasonable Doubt (2021) 607

that, at least for this specification of preferences, the Bayes decision rule is relatively robust to ambiguity over the nature of the true DGP. As shown in Appendix Figure A.II, however, had we considered a more aggressive Bayesian benchmark (e.g., a 40% threshold), the minimax and Bayes rules would have departed more substantially, as the maximum risk can greatly exceed the Bayes risk.

That the Bayes and minimax thresholds are nearer each other for higher nominal thresholds is to be expected, as our risk function must approach zero as the posterior cutoff approaches one. In practice, state and regulatory agencies are likely to exhibit preferences requiring relatively high posterior thresholds that, as in our example, make the bounds on the risk function relatively narrow over the relevant range of decision rules. The concordance of maximum and Bayes risk over this range suggests that flexible parametric models such as the mixed logit specification employed in Section 8 may serve as a useful hueristic for decisionmaking by such entities.

10 Conclusion

Correspondence studies are powerful tools that have been extensively used to detect market-level averages of discriminatory behavior. Revisiting three such studies, we find tremendous heterogeneity across jobs in the degree of discrimination. This heterogeneity presents authorities charged with enforcing anti-discrimination laws with a difficult inferential task. Our analysis suggests that when ensemble evidence is used, sending 10 applications per job enables accurate detection of a non-trivial share of discriminatory employers. This finding opens the possibility that discrimination can be monitored – perhaps in real time – at the employer level.

Our results also provide a number of methodological lessons regarding the design and analysis of correspondence studies, and of experimental ensembles more generally. First, we demonstrate that indirect evidence can serve as a valuable supplement to direct evidence even when heterogeneity distributions are not point identified. Using a few moments of the callback rate distribution in conjunction with only four applications per job, we derived informative lower bounds on the share of jobs engaged in illegal discrimination in the NPRS experiment. In the AGCV study, which sent eight applications to each job, we deduced informative lower bound rates of discrimination against men and women separately.

Second, our results highlight that the appropriate use of indirect evidence depends critically on the objectives of the investigator, formalized in our framework by the loss function of a hypothetical regulator. While in point identified settings it is straightforward to characterize the tradeoffs presented by different decision rules, partial identification of heterogeneity distributions tends to undermine identifiability of this tradeoff itself. In our setting acknowledging the ambiguity stemming from partial identification turns out to lead to only slightly more conservative decisions with a realistic loss function. An important topic for future research is the extent to which the policy implications of recent econometric evaluations of teachers, schools, hospitals, and neighborhoods (e.g., Chetty et al., 2014b; Angrist et al., 2017; Hull, 2018; Chetty and Hendren, 2018; Chetty et al.,

2018) vary with alternative notions of risk.

References

7th Circuit Court of Appeals (2006): "EEOC v. Target Corp." 460 (F. 3d), 946.

Altonji, J. G. and R. M. Blank (1999): "Race and gender in the labor market," in *Handbook of Labor Economics*, ed. by O. C. Ashenfelter and D. Card, Elsevier, vol. 3C, chap. 48, 3143–3259.

Angrist, J. D., P. D. Hull, P. A. Pathak, and C. R. Walters (2017): "Leveraging lotteries for school value-added: testing and estimation," *Quarterly Journal of Economics*, 132, 871–919.

Arceo-Gomez, E. O. and R. M. Campos-Vasquez (2014): "Race and marriage in the labor market: a discrimination correspondence study in a developing country," *American Economic Review: Papers & Proceedings*, 104, 376–380.

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

Benjamini, Y. and Y. Hochberg (1995): "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society*, 57, 289–300.

Bertrand, M. and E. Duflo (2017): "Field experiments on discrimination," in *Handbook of Field Experiments*, ed. by E. Duflo and A. Banerjee, Elsevier, vol. 1.

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

Brown, L. D. (2008): "In-season prediction of batting averages: A field test of empirical Bayes and Bayes methodologies," *The Annals of Applied Statistics*, 2, 113–152.

Chernozhukov, V., W. K. Newey, and A. Santos (2015): "Constrained conditional moment restriction models," *arXiv preprint arXiv:1509.06311*.

Chetty, R., J. N. Friedman, N. Hendren, M. R. Jones, and S. R. Porter (2018): "The opportunity atlas: mapping the childhood roots of social mobility," Working paper.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014a): "Measuring the impact of teachers I: evaluating bias in teacher value-added estimates," *American Economic Review*, 104, 2593–2563.

——— (2014b): "Measuring the impact of teachers II: teacher value-added and student outcomes in adulthood," *American Economic Review*, 104, 2633–2679.

Chetty, R. and N. Hendren (2018): "Impacts of neighborhoods on intergenerational mobility II: county-level estimates," *Quarterly Journal of Economics*, 133, 1163–1228.

Chetverikov, D., A. Santos, and A. M. Shaikh (2018): "The econometrics of shape restrictions," *Annual Review of Economics*, 10, 31–63.

Efron, B. (2010): "The future of indirect evidence," *Statistical Science*, 25, 145–157.

——— (2012): *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, vol. 1, Cambridge University Press.

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

Efron, B. and C. Morris (1975): "Data analysis using Stein's estimator and its generalizations," *Journal of the American Statistical Association*, 70, 311–319.

Efron, B., R. Tibshirani, J. D. Storey, and V. Tusher (2001): "Empirical Bayes analysis of a microarray experiment," *Journal of the American Statistical Association*, 96, 1151–1160.

Fang, Z. and A. Santos (2018): "Inference on directionally differentiable functions," *The Review of Economic Studies*, 86, 377–412.

Farber, H. S., D. Silverman, and T. von Wachter (2016): "Determinants of callbacks to job applications: an audit study," *American Economic Review: Papers & Proceedings*, 106, 314–318.

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

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

Halevy, Y., D. Persitz, and L. Zrill (2018): "Parametric recoverability of preferences," *Journal of Political Economy*, 126, 1558–1593.

Hong, H. and J. Li (forthcoming): "The numerical bootstrap," *Annals of Statistics*.

Hull, P. D. (2018): "Estimating hospital quality with quasi-experimental data," Working paper.

Levine, D. I., M. W. Toffel, and M. S. Johnson (2012): "Randomized government safety inspections reduce worker injuries with no detectable job loss," *Science*, 336, 907–911.

Manski, C. F. (2000): "Identification problems and decisions under ambiguity: Empirical analysis of treatment response and normative analysis of treatment choice," *Journal of Econometrics*, 95, 415–442.

Noubiap, R. F., W. Seidel, et al. (2001): "An algorithm for calculating Γ-minimax decision rules under generalized moment conditions," *The Annals of Statistics*, 29, 1094–1116.

Nunley, J. M., A. Pugh, N. Romero, and R. A. Seals (2015): "Racial discrimination in the labor market for recent college graduates: evidence from a field experiment," *B.E. Journal of Economic Analysis and Policy*, 15, 1093–1125.

Pearson, K. (1900): "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *Philosophical Magazine, Series 5*, 50, 157–175.

Rubin, D. B. (1980): "Randomization analysis of experimental data: the Fisher Randomization test comment," *Journal of the American Statistical Association*, 75, 591–593.

Savage, L. J. (1951): "The theory of statistical decision," *Journal of the American Statistical association*, 46, 55–67.

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

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

US Equal Employment Opportunity Commission (2016): "Advancing opportunity: a review of the systematic program of the US Equal Employment Opportunity Commission," Technical report.

Wald, A. (1945): "Statistical decision functions which minimize the maximum risk," *Annals of Mathematics*, 46, 265–280.

Appendix A: Connections to Large Scale Testing

The problem of detecting individual discriminators based on correspondence evidence is closely related to the literature on large scale testing, which is concerned with deciding which hypotheses to reject based upon the results of a very large number of tests (Efron, 2012 provides a review). A seminal contribution to this literature comes from Benjamini and Hochberg (1995), who proposed controlling the False Discovery Rate (FDR): the expected share of rejected null hypotheses that are true. We next show that a decision rule based on the posterior probability $\pi(c_w, c_b)$ will control an analogue of the FDR, while a decision rule based on classical hypothesis testing will not.

As in Section 9, let $\delta : \{0, ..., L_w\} \times \{0, ..., L_b\} \to \{0, 1\}$ represent an auditing rule that maps the evidence vector (C_{jw}, C_{jb}) to a binary investigation decision. Letting $N_J \equiv \sum_{j=1}^{J} \delta(C_{jw}, C_{jb})$ denote the total number of investigations in a sample of J jobs, we can define the *Positive False Discovery Rate* (Storey, 2003) as: $pFDR_J = \mathbb{E}\left[N_J^{-1} \sum_{j=1}^{J} \delta(C_{jw}, C_{jb})(1 - D_j) | N_J \geq 1\right]$. In words, $pFDR_J$ gives the proportion of investigated jobs that are not discriminating, conditional on at least one investigation taking place. The following Lemma establishes that a posterior cutoff decision rule controls $pFDR_J$.

Lemma 4 ($pFDR_J$ Control)**.** *If* $\delta\left(C_{jw}, C_{jb}\right) = 1\left\{\pi(C_{jw}, C_{jb}) > \bar{p}\right\}$ *then* $pFDR_J \leq 1 - \bar{p}$.

Proof. Storey (2003, Theorem 1) showed that $pFDR_J = \Pr(D_j = 0 | \delta(C_{jw}, C_{jb}) = 1)$ for any deterministic decision rule $\delta(\cdot)$ obeying $\Pr\left(\delta(C_{jw}, C_{jb}) = 1\right) > 0$. Then the posterior cutoff rule $\delta\left(C_{jw}, C_{jb}\right) = 1\left\{\pi(C_{jw}, C_{jb}) > \bar{p}\right\}$ yields

$$\begin{aligned} pFDR_J &= \Pr\left(D_j = 0 | \pi(C_{jw}, C_{jb}) > \bar{p}\right) \\ &\leq \Pr\left(D_j = 0 | \pi(C_{jw}, C_{jb}) = \bar{p}\right) = 1 - \bar{p}. \end{aligned}$$

□

By contrast, consider an alternative decision rule $\delta^{\dagger}\left(C_{jw}, C_{jb}\right)$ based on a classical hypothesis test that controls size at a fixed level $\tilde{\alpha} < 1$. To simplify exposition, suppose that the test is pivotal under the null of non-discrimination so that

$$\Pr\left(\delta^{\dagger}\left(C_{jw}, C_{jb}\right) = 1 | p_{jw} = p, p_{jb} = p\right) = \tilde{\alpha}, \quad \forall p \in [0, 1].$$

We can write the resulting $pFDR_J$ of this rule

$$\begin{aligned} \Pr\left(D_j = 0 | \delta^{\dagger}\left(C_{jw}, C_{jb}\right) = 1\right) &= \frac{\Pr\left(\delta^{\dagger}\left(C_{jw}, C_{jb}\right) = 1 | D_j = 0\right)(1 - \bar{\pi})}{\Pr\left(\delta^{\dagger}\left(C_{jw}, C_{jb}\right) = 1 | D_j = 0\right)(1 - \bar{\pi}) + \Pr\left(\delta^{\dagger}\left(C_{jw}, C_{jb}\right) = 1 | D_j = 1\right)\bar{\pi}} \\ &\geq \frac{\tilde{\alpha}(1 - \bar{\pi})}{\tilde{\alpha}(1 - \bar{\pi}) + \bar{\pi}}. \end{aligned}$$

To see that $\delta^{\dagger}\left(C_{jw}, C_{jb}\right)$ fails to control $pFDR_J$, note that $\lim_{\bar{\pi} \downarrow 0} \frac{\tilde{\alpha}(1-\bar{\pi})}{\tilde{\alpha}(1-\bar{\pi})+\bar{\pi}} = 1$. That is, when almost no jobs are discriminating, classical hypothesis testing will result in the vast majority of

investigations being false accusations.

The *False Discovery Rate* of Benjamini and Hochberg (1995) can be written $FDR_J = pFDR_J \times \Pr(N_J \geq 1)$. Because $\Pr(N_J \geq 1) \leq 1$, Lemma 4 implies that the posterior cutoff rule also controls FDR_J.

Appendix B: Proof of Lemma 2

By the law of total probability the share of jobs calling c_w white and $t - c_w$ black applications among those calling t total can be written:

$$\bar{f}_t(c_w) = (1 - \bar{\pi}_t)\bar{f}_t^0(c_w) + \bar{\pi}_t \bar{f}_t^1(c_w),$$

where $\bar{f}_t^d(c_w) = \Pr(C_{jw} = c_w | C_{jw} + C_{jb} = t, D_j = d)$ for $d \in \{0, 1\}$. Since $\bar{f}_t^1(c_w) \in [0, 1]$ we have

$$\bar{f}_t(c_w) \geq (1 - \bar{\pi}_t)\bar{f}_t^0(c_w), \;\; \bar{f}_t(c_w) \leq (1 - \bar{\pi}_t)\bar{f}_t^0(c_w) + \bar{\pi}_t,$$

which implies

$$\bar{\pi}_t \geq \max\left\{ \tfrac{\bar{f}_t^0(c_w) - \bar{f}_t(c_w)}{\bar{f}_t^0(c_w)}, \tfrac{\bar{f}_t(c_w) - \bar{f}_t^0(c_w)}{1 - \bar{f}_t^0(c_w)} \right\}.$$

Taking the maximum of these lower bounds over $c_w \in \{0, ..., t\}$ yields the bound on $\bar{\pi}_t$ in part i) of Lemma 2.

By Bayes' rule the share of discriminators among jobs calling c_w white and $t - c_w$ black applications is given by:

$$\pi(c_w, t - c_w) = 1 - \frac{\bar{f}_t^0(c_w)(1 - \bar{\pi}_t)}{\bar{f}_t(c_w)}.$$

Plugging the bound on $\bar{\pi}_t$ from part i) of the Lemma into this expression gives the bound on $\pi(c_w, t - c_w)$ in part ii).

Appendix C: Discretization of G and Linear Programming Bounds

To compute the solution to the problem in (5), we approximate the CDF $G(p_w, p_b)$ with the discrete distribution

$$G_K(p_w, p_b) \;\; = \;\; \sum_{k=1}^{K} \sum_{s=1}^{K} \eta_{ks} 1\left\{ p_w \leq \varrho(k, s), p_b \leq \varrho(s, k) \right\},$$

where the $\{\eta_{ks}\}_{k=1,s=1}^{K,K}$ are probability masses and $\{\varrho(k, s), \varrho(s, k)\}_{k=1,s=1}^{K,K}$ comprise a set of mass point coordinates generated by the function

$$\varrho(k, s) = \underbrace{\frac{\min\{k, s\} - 1}{K}}_{\text{diagonal}} + \underbrace{\frac{\max\{0, k - s\}^2}{K(1 + K - y)}}_{\text{off-diagonal}}.$$

This discretization scheme can be visualized as a two-dimensional grid containing K^2 elements. The diagonal entries on the grid represent jobs where no discrimination is present. The first term above ensures the mass points are equally spaced along the diagonal from $(0,0)$ to $\left(\frac{K-1}{K}, \frac{K-1}{K}\right)$. The second term spaces off diagonal points quadratically according to their distance from the diagonal in order to accomodate jobs with very low levels of discrimination while economizing on the number of grid points. We use a spacing scheme that places more points near the diagonal because we are particularly interested in the mass exactly on the diagonal. Note that $\lim_{K\to\infty} \varrho(K,s) = 1$, ensuring the grid asymptotically spans the unit square.

With this notation, the constraints in (6) can be written:

$$\bar{f}(c_w, c_b) = \binom{L_w}{c_w}\binom{L_b}{c_b}\sum_{k=1}^{K}\sum_{s=1}^{K}\eta_{ks}\varrho(k,s)^{c_w}(1-\varrho(k,s))^{L_w-c_w}\varrho(s,k)^{c_b}(1-\varrho(s,k))^{L_b-c_b}, \quad (10)$$

for $c_w = (1,...,L_w)$ and $c_b = (0,...,L_b)$. Hence, our composite discretized optimization problem is to

$$\min_{\{\eta_{ks}\}} 1 - \frac{\binom{L}{t}}{\sum_{(c'_w,c'_b):c'_w+c'_b=t}\bar{f}\left(c'_w,c'_b\right)}\sum_{k=1}^{K}\eta_{kk}\varrho(k,k)^t(1-\varrho(k,k))^{L-t},$$

subject to (10) and

$$\sum_{k=1}^{K}\sum_{s=1}^{K}\eta_{ks} = 1, \quad \eta_{ks} \geq 0,$$

for $k = 1,...,K$ and $m = 1,...,K$. We solve this problem numerically using the Gurobi software package. Because setting K too low will tend to yield artificially tight bounds, we set $K = 900$ in all bound computation steps, which yields $(900)^2 = 810,000$ distinct mass points.

Appendix Table A.IV reports linear programming bounds for various choices of K. As expected, the bounds stabilize with a sufficiently large K, and the quadratic spacing described above produces more accurate results than an equally-spaced grid: we obtain similar estimates for a quadratic grid with 300^2 grid points and a rectangular grid with 900^2 points.

Appendix D: Shape Constrained GMM

To accomodate the Nunley et al. (2015) study which employs multiple application designs, we introduce the variable $L_j = (L_{jw}, L_{jb})$ which gives the number of white and black applications sent to job j. Collecting the design-specific callback probabilities $\{\Pr(C_{jw} = c_w, C_{jb} = c_b | L_j = l)\}_{c_w,c_b}$ into the vector $\bar{f}_l$, our model relates these probabilities to moments of the callback distribution via the linear system $\bar{f}_l = B_l\mu$, for B_l a fixed matrix of binomial coefficients. Letting $\bar{f}$ denote the vector formed by "stacking" the $\left\{\bar{f}_l\right\}$ across designs in an experiment, we write $\bar{f} = B\mu$. Let η be a $K^2 \times 1$ vector comprised of the probability masses $\{\eta_{ks}\}_{k=1,s=1}^{K,K}$ (see Appendix C). For GMM estimation we set $K = 150$ (larger values yield very similar results). From (3), we can write $\mu = M\eta$ where M is a $dim(\mu) \times K^2$ matrix comprised of entries with typical element $\varrho(k,s)^m\varrho(s,k)^n$. We

From callbacks to labor markets - Kline & Walters, Reasonable Doubt (2021) 614

then have the moment restriction $\bar{f} = BM\eta$.

Let $\tilde{f}$ denote the vector of empirical call back probabilities with typical element:

$$\frac{J^{-1}\sum_{j=1}^{J} 1\{C_{jw}=c_w, C_{jb}=c_b, L_j=l\}}{J^{-1}\sum_{j=1}^{J} 1\{L_j=l\}}.$$

Our shape constrained GMM estimator of η can be written as the solution to the following quadratic programming problem:

$$\hat{\eta} = \arg\inf_{\eta} (\tilde{f} - BM\eta)' W (\tilde{f} - BM\eta) \tag{11}$$

$$\text{s.t.} \quad \eta \geq 0, \ \mathbf{1}'\eta = 1,$$

where W is a fixed weighting matrix. Note that because $G(\cdot,\cdot)$ is not identified, there are many possible solutions $\hat{\eta}$ to this problem, but these solutions will all yield the same values of $BM\hat{\eta}$. Our shape constrained estimate of the moments is $\hat{\mu} = M\hat{\eta}$ while our estimator of the callback probabilities is $\hat{\bar{f}} = BM\hat{\eta}$. We follow a two-step procedure, solving (11) with diagonal weights proportional to the number of jobs used in the application design and then choosing $W = \hat{\Sigma}^{-1}$ where $\hat{\Sigma} = \text{diag}\left(\hat{\bar{f}}^{(1)}\right) - \hat{\bar{f}}^{(1)}\hat{\bar{f}}^{(1)\prime}$ is an estimate of the variance-covariance matrix of the callback frequencies implied by the first step shape-constrained callback probability estimates $\hat{\bar{f}}^{(1)}$.

Hong and Li (forthcoming) standard errors

Standard errors on the moment estimates $\hat{\mu}$ are computed via the numerical bootstrap procedure of Hong and Li (forthcoming) using a step size of $J^{-1/3}$ (we found qualitatively similar results with a step size of $J^{-1/4}$). Our implementation of the numerical bootstrap proceeds as follows: the bootstrap analogue μ^* of $\hat{\mu}$ solves the quadratic programming problem in (11) where $\tilde{f}$ has been replaced by $\left(\tilde{f} + J^{-1/3} f^*\right)$. The bootstrap probabilities f^* have typical element:

$$J^{1/2}\left(\frac{\sum_{j=1}^{J}\omega_j^* 1\{C_{jw}=c_w, C_{jb}=c_b, L_j=l\}}{\sum_{j=1}^{J}\omega_j^* 1\{L_j=l\}} - \frac{\sum_{j=1}^{J} 1\{C_{jw}=c_w, C_{jb}=c_b, L_j=l\}}{\sum_{j=1}^{J} 1\{L_j=l\}}\right),$$

where $\left\{\omega_j^*\right\}_{j=1}^{J}$ are a set of iid draws from an exponential distribution with mean and variance one. For any function $\phi(\hat{\mu})$ of the moment estimates $\hat{\mu}$ reported, we use as our standard error estimate the standard deviation across bootstrap replications of $J^{-1/3}\left[\phi(\mu^*) - \phi(\hat{\mu})\right]$.

Chernozhukov et al. (2015) goodness of fit test

To formally test whether there exists an η in the K^2 dimensional probability simplex such that $f = BM\eta$ holds, we rely on the procedure of Chernozhukov et al. (2015). Our test statistic (the "J-test") can be written:

$$T_n = \inf_{\eta} (\tilde{f} - BM\eta)' \hat{\Sigma}^{-1} (\tilde{f} - BM\eta)$$

$$\text{s.t.} \quad \eta \geq 0,\ \mathbf{1}'\eta = 1.$$

Letting $\mathbb{F}^* = f^* - \tilde{f}$ denote the (centered) bootstrap analogue of the callback frequencies $\tilde{f}$ and W^* a corresponding bootstrap weighting matrix, our bootstrap test statistic takes the form:

$$T_n^* = \inf_{\eta,h} (\mathbb{F}^* - BMh)'W^*(\mathbb{F}^* - BMh) \tag{12}$$

$$\text{s.t.} \quad (\tilde{f} - BM\eta)'W(\tilde{f} - BM\eta) = T_n,\ \eta \geq 0,\ \mathbf{1}'\eta = 1,\ h \geq -\eta,\ \mathbf{1}'h = 0.$$

As in the full sample problem, we conduct a two-step GMM procedure in each bootstrap replication, setting $W^* = \left[\text{diag}(BM\eta^{(1)*}) - (BM\eta^{(1)*})(BM\eta^{(1)*})'\right]^{-1}$ where $\eta^{(1)*}$ is a first-step diagonally weighted estimate of the probabilities in the bootstrap sample. The goodness of fit p-value reported is the share of bootstrap samples for which $T_n^* > T_n$.

To simplify computation of (12), we re-formulate the problem in two ways. First, we define primary and auxilliary vectors of errors for each moment condition. Letting $\xi_h = \mathbb{F}^* - BMh$ and $\xi_\eta = \tilde{f} - BM\eta$, the problem can be re-posed as:

$$T_n^* = \inf_{\xi_h,\xi_\eta} \xi_h' W^* \xi_h,$$

$$\text{s.t.} \quad \xi_\eta' W \xi_\eta = T_n,\ BMh + \xi_h = \mathbb{F}^*,\ BM\eta + \xi_\eta = \tilde{f},\ \mathbf{1}'h = 0,\ \mathbf{1}'\eta = 1,\ h \geq -\eta,\ \eta \geq 0.$$

Now letting $h^+ = h + \eta$, we can further rewrite the problem as:

$$T_n^* = \inf_{\xi_h,\xi_\eta} \xi_h' W^* \xi_h,$$

$$\text{s.t.} \quad \xi_\eta' W \xi_\eta = T_n,\ BMh^+ + \xi_h + \xi_\eta = \mathbb{F}^*,\ BM\eta + \xi_\eta = \tilde{f},\ \mathbf{1}'h^+ = 1,\ \mathbf{1}'\eta = 1,\ h^+ \geq 0,\ \eta \geq 0.$$

Note that this final representation replaces a $K^2 \times K^2 + 1$ (inequality) constraint matrix encoding $\xi_h \geq -\xi_\eta$ and $\xi_\eta \geq 0$ with a $2K^2 \times 1$ vector encoding $h^+ \geq 0$ and $\eta \geq 0$. Because this problem still involves a quadratic constraint in ξ_η, we make use of Gurobi's Second Order Cone Programming (SOCP) solver to obtain a solution.

Appendix E: Computing Maximum Risk

We approximate $G\left(p_w^H, p_w^L, p_b^H, p_b^L\right)$ with the discretized distribution

$$G_K\left(p_w^H, p_w^L, p_b^H, p_b^L\right) = \sum_{k=1}^{K}\sum_{s=1}^{K}\sum_{k'=1}^{K}\sum_{s'=1}^{K} \eta_{ksk's'} 1\left\{p_w^H \leq \varrho\left(k,s\right), p_w^L \leq \varrho\left(k',s'\right), p_b^H \leq \varrho\left(s,k\right), p_b^L \leq \varrho\left(s',k'\right)\right\},$$

which has K^4 mass points. In practice, we choose $K = 30$, which yields the same number of points as the approximation described in Appendix C.

Generalizing the notation of Appendix D, let the vector $L_j = \left(L_{jw}^H, L_{jw}^L, L_{jb}^H, L_{jb}^L\right)$ record the number of high quality and low quality applications of each race sent to job j and let $C_j = \left(C_{jw}^H, C_{jw}^L, C_{jb}^H, C_{jb}^L\right)$ record the corresponding numbers of callbacks. The space of auditing rules we consider is of the form $\delta\left(C_j, L_j, q\right) = 1\left\{\mathcal{P}\left(C_j, L_j, G_{logit}\right) > q\right\}$.With this notation, we can write the risk function

$$\mathcal{R}(q) = \sum_{l \in \mathscr{A}_1} w_l \mathbb{E}\left[\delta(C_j, l, q)\left\{\kappa - \Lambda\left(\sum_{x \in \{H,L)} \frac{\Lambda^{-1}(p_{wj}^x) - \Lambda^{-1}(p_{bj}^x)}{2}\right)\right\} \mid L_j = l\right],$$

where $\mathscr{A}_1$ is the set of all $2^5 = 36$ binary quality permutations possible in a design with 5 white and 5 black applications and $w_l = \binom{5}{l_w^H}\binom{5}{l_b^H}(1/2)^{10}$ is the set of weights that arise when quality is assigned at random within race.

To further evaluate the above risk expression we can write:

$$\mathbb{E}\left[\delta(C_j, l, q)\left\{\kappa - \Lambda\left(\sum_{x \in \{H,L)} \frac{\Lambda^{-1}(p_{wj}^x) - \Lambda^{-1}(p_{bj}^x)}{2}\right)\right\} \mid L_j = l\right] =$$

$$\sum_{c_w^H=0}^{a_{jw}^H} \sum_{c_b^H=0}^{a_{jb}^H} \sum_{c_w^L=0}^{a_{jw}^L} \sum_{c_b^L=0}^{a_{jb}^L} \sum_{k=1}^{K} \sum_{s=1}^{K} \sum_{k' \geq k}^{K} \sum_{s' \geq s}^{K} \delta(c, l, q) \eta_{ksk's'} \binom{l_w^H}{c_w^H}\binom{l_b^H}{c_b^H}\binom{l_w^L}{c_w^L}\binom{l_b^L}{c_b^L}$$

$$\times \varrho\left(k,s\right)^{c_w^H}\left(1-\varrho\left(k,s\right)\right)^{l_w^H - c_w^H} \varrho\left(s,k\right)^{c_b^H}\left(1-\varrho\left(s,k\right)\right)^{l_b^H - c_b^H}$$

$$\times \varrho\left(k',s'\right)^{c_w^L}\left(1-\varrho\left(k',l=s'\right)\right)^{l_w^L - c_w^L} \varrho\left(s',k'\right)^{c_b^L}\left(1-\varrho\left(s',k'\right)\right)^{l_b^L - c_b^L}$$

$$\times \left\{\kappa - \Lambda\left(\frac{\Lambda^{-1}(\varrho(k,s)) - \Lambda^{-1}(\varrho(s,k))}{2} + \frac{\Lambda^{-1}(\varrho(k',s')) - \Lambda^{-1}(\varrho(s',k'))}{2}\right)\right\}.$$

Using this expression, maximal risk can therefore be written as the solution to the following linear programming problem:

$$\mathcal{R}^m(q) = \max_{\{\eta_{ksk's'}\}} \sum_{l \in \mathscr{A}_1} w_l \mathbb{E}\left[\delta(C_j, l, q)\left\{\kappa - \Lambda\left(\sum_{x \in \{H,L)} \frac{\Lambda^{-1}(p_{wj}^x) - \Lambda^{-1}(p_{bj}^x)}{2}\right)\right\} \mid L_j = l\right]$$

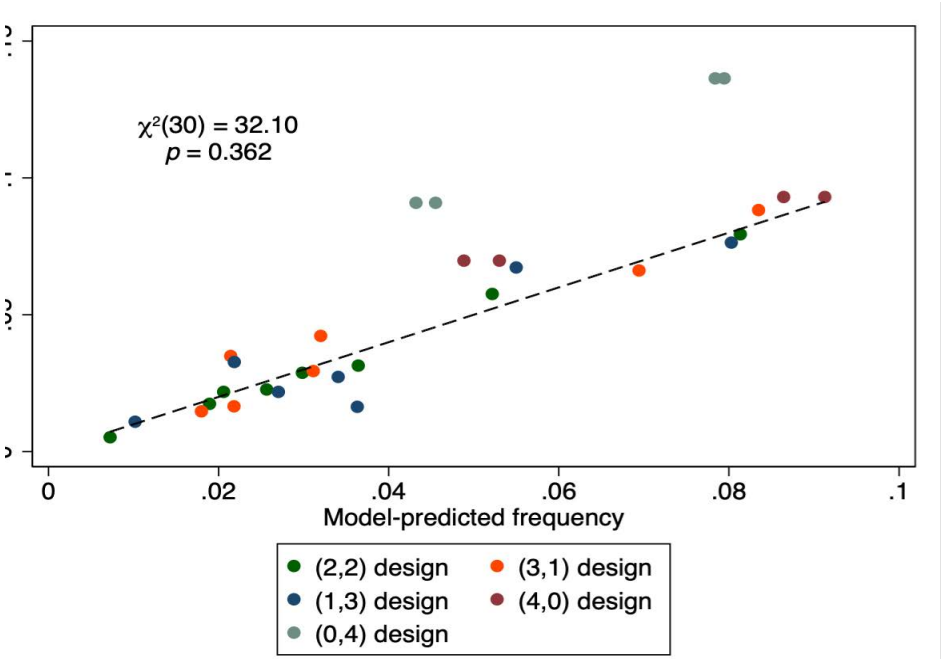
subject to the constraint that the $\eta_{klk'l'}$ are non-negative and sum to one and that the following moment restrictions hold:

$$\begin{aligned}\Pr\left(C_j = c | L_j = l\right) &= \binom{l_w^H}{c_w^H}\binom{l_b^H}{c_b^H}\binom{l_w^L}{c_w^L}\binom{l_b^L}{c_b^L} \sum_{k=1}^{K}\sum_{s=1}^{K}\sum_{k'=1}^{K}\sum_{s'=1}^{K} \eta_{ksk's'} \\ &\times \varrho\left(k,s\right)^{c_w^H}\left(1-\varrho\left(k,s\right)\right)^{l_w^H - c_w^H} \varrho\left(s,k\right)^{c_b^H}\left(1-\varrho\left(s,k\right)\right)^{l_b^H - c_b^H} \\ &\times \varrho\left(k',s'\right)^{c_w^L}\left(1-\varrho\left(k',s'\right)\right)^{l_w^L - c_w^L} \varrho\left(s',k'\right)^{c_b^L}\left(1-\varrho\left(s',k'\right)\right)^{l_b^L - c_b^L}.\end{aligned}$$

We impose these restrictions for the following set of designs, all of which are present in the Nunley et al. (2015) experiment: $\mathscr{A}_2 = \{(2,0,2,0),(2,0,0,2),(0,2,2,0),(0,2,0,2)\}$.To operationalize

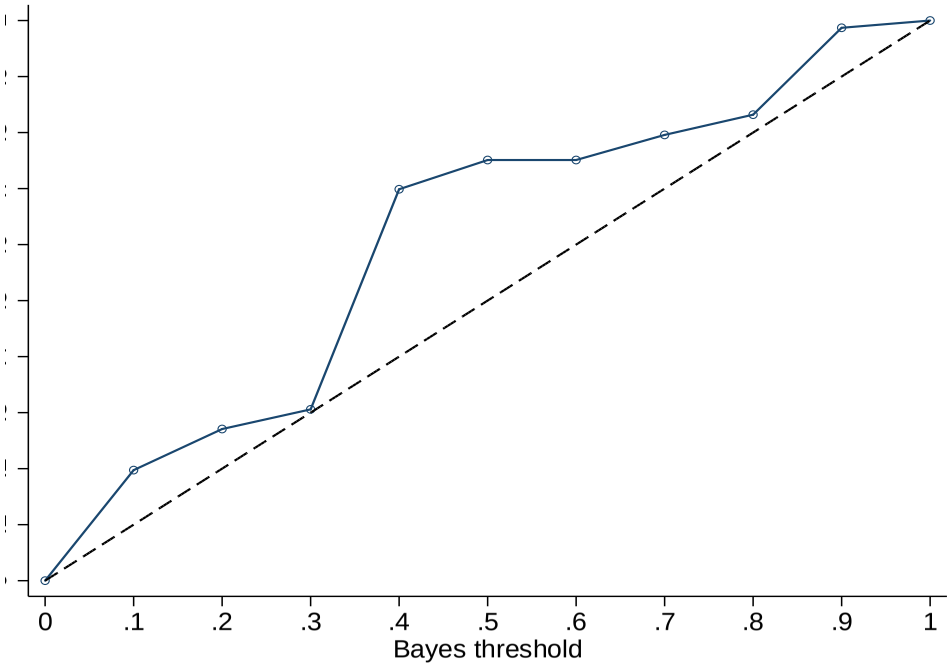
these constraints, we replace the unknown cell probabilities $\Pr(C_j = c|L_j = l)$ for all c and l in $\mathscr{A}_2$ with their predictions under the logit model reported in column 2 of Table V. Using the logit predictions serves as a form of smoothing that allows us to avoid problems that arise with small cells when considering quality variation due to covariates.

Figure A.I: Mixed logit model fit



This figure compares mixed logit predicted frequencies for callback events in the Nunley et al. (2015) th corresponding empirical frequencies. The horizontal axis plots model-predicted probabilities for ssible combination of white and black callback counts (excluding zero total callbacks), separately by ental design. Model predictions are calculated by simulating the logit model in column (2) of Table X times for each job in the Nunley et al. data set. The vertical axis plots the observed frequency of each Green dots show frequencies for a design with two white and two black applications, while orange, d, and grey points show frequencies for designs with 3 white and 1 black, 1 white and three black, 4 nd zero black, and 0 white and 4 black applications, respectively. The dashed line is the 45-degree line. i-squared statistic and *p* -value come from a test that all model-predicted and empirical frequencies treating the model predictions as fixed.

Figure A.II: Bayes and minimax investigation thresholds



his figure compares Bayes and minimax decisions for various values of the investigation cost parameter κ. ontal axis displays the posterior investigation threshold for a Bayes regulator for each value of κ, and the xis shows the corresponding threshold for a minimax regulator. The dashed line is the 45 degree line.

Table A.I: Moments of callback rate distribution, BM data

Moment	No constraints (1)	Shape constraints (2)
$E[p_w]$	0.094	0.094
	(0.006)	(0.007)
$E[p_b]$	0.063	0.063
	(0.006)	(0.006)
$E[(p_w - E[p_w])^2]$	0.040	0.040
	(0.005)	(0.005)
$E[(p_b - E[p_b])^2]$	0.023	0.023
	(0.004)	(0.004)
$E[(p_w - E[p_w])(p_b - E[p_b])]$	0.028	0.028
	(0.004)	(0.003)
$E[(p_w - E[p_w])^2(p_b - E[p_b])]$	0.015	0.014
	(0.003)	(0.002)
$E[(p_w - E[p_w])(p_b - E[p_b])^2]$	0.023	0.012
	(0.003)	(0.002)
$E[(p_w - E[p_w])^2(p_b - E[p_b])^2]$	0.010	0.010
	(0.003)	(0.002)
	J-statistic:	0.0
	P-value:	1.00
Sample size	1,112	

Notes: This table reports generalized method of moments (GMM) estimates of moments of the joint distribution of job-specific white and black callback rates in the Bertrand and Mullainathan (2004) data. Estimates in column (2) come from a shape-constrained GMM procedure imposing that the moments are consistent with a well-defined probability distribution. The J-statistic is the minimized shape-constrained GMM criterion function. The p-value come from a bootstrap test of the hypothesis that the model restrictions are satisfied.

Table A.II: Moments of callback rate distribution, NPRS data

Moment	Design-specific estimates: (2,2) design (1)	(3,1) design (2)	(1,3) design (3)	P-value (4)	Combined estimates (5)
$E[p_w]$	0.174	0.199	0.142	0.027	0.177
	(0.010)	(0.025)	(0.015)		(0.007)
$E[p_b]$	0.148	0.149	0.157	0.854	0.153
	(0.010)	(0.015)	(0.013)		(0.007)
$E[(p_w - E[p_w])^2]$	0.089	0.108	-	0.097	0.095
	(0.007)	(0.009)			(0.005)
$E[(p_b - E[p_b])^2]$	0.085	-	0.083	0.857	0.084
	(0.007)		(0.008)		(0.005)
$E[(p_w - E[p_w])(p_b - E[p_b])]$	0.083	0.084	0.080	0.926	0.084
	(0.006)	(0.009)	(0.009)		(0.004)
$E[(p_w - E[p_w])^3]$	-	0.051	-		0.106
		(0.008)			(0.007)
$E[(p_b - E[p_b])^3]$	-	-	0.044		0.092
			(0.007)		(0.006)
$E[(p_w - E[p_w])^2(p_b - E[p_b])]$	0.044	0.043	-	0.875	0.040
	(0.004)	(0.007)			(0.002)
$E[(p_w - E[p_w])(p_b - E[p_b])^2]$	0.047	-	0.045	0.819	0.042
	(0.005)		(0.007)		(0.002)
$E[(p_w - E[p_w])^3(p_b - E[p_b])]$	-	0.034	-	-	0.035
		(0.005)			(0.002)
$E[(p_w - E[p_w])(p_b - E[p_b])^3]$	-	-	0.037	-	0.037
			(0.006)		(0.002)
$E[(p_w - E[p_w])^2(p_b - E[p_b])^2]$	0.036	-	-	-	0.038
	(0.004)				(0.002)
				J-statistic:	23.1
				P-value:	0.190
Sample size	1,146	544	550		2,240

Notes: This table reports generalized method of moments (GMM) estimates of moments of the joint distribution of job-specific white and black callback rates in the Nunley et al. (2015) data. Columns (1), (2), and (3) show estimates based on jobs that received 2 white and 2 black, 3 white and 1 black, and 1 white and 3 black applications, respectively. Column (4) shows p-values from tests that the moments are the same in each design. Estimates in column (5) come from a shape-constrained GMM procedure imposing that the moments are consistent with a well-defined probability distribution. The J-statistic is the minimized shape-constrained GMM criterion function. The p-value come from a bootstrap test of the hypothesis that the model restrictions are satisfied.

Table A.III: Moments of callback rate distribution, AGCV data

Moment	No constraints (1)	Shape constraints (2)	Moment	No constraints (3)	Shape constraints (4)
$E[p_f]$	0.136 (0.010)	0.137 (0.010)	$E\left[(p_f - E[p_f])^4\right]$	0.024 (0.004)	0.026 (0.003)
$E[p_m]$	0.103 (0.009)	0.109 (0.009)	$E[(p_m - E[p_m])^4]$	0.019 (0.004)	0.023 (0.003)
$E\left[(p_f - E[p_f])^2\right]$	0.066 (0.006)	0.066 (0.006)	$E\left[(p_f - E[p_f])^4 (p_m - E[p_m])\right]$	0.012 (0.003)	0.012 (0.002)
$E[(p_m - E[p_m])^2]$	0.047 (0.005)	0.052 (0.006)	$E\left[(p_f - E[p_f])(p_m - E[p_m])^4\right]$	0.013 (0.003)	0.013 (0.002)
$E\left[(p_f - E[p_f])(p_m - E[p_m])\right]$	0.043 (0.005)	0.043 (0.004)	$E\left[(p_f - E[p_f])^3 (p_m - E[p_m])^2\right]$	0.012 (0.003)	0.011 (0.002)
$E\left[(p_f - E[p_f])^3\right]$	0.032 (0.005)	0.064 (0.007)	$E\left[(p_f - \mu_f)^2 (p_m - E[p_m])^3\right]$	0.012 (0.003)	0.013 (0.002)
$E[(p_m - E[p_m])^3]$	0.025 (0.005)	0.048 (0.007)	$E\left[(p_f - E[p_f])^4 (p_m - E[p_m])^2\right]$	0.010 (0.002)	0.010 (0.002)
$E\left[(p_f - E[p_f])^2 (p_m - E[p_m])\right]$	0.021 (0.004)	0.018 (0.003)	$E\left[(p_f - E[p_f])^2 (p_m - E[p_m])^4\right]$	0.010 (0.002)	0.010 (0.002)
$E\left[(p_f - E[p_f])(p_m - E[p_m])^2\right]$	0.022 (0.004)	0.020 (0.003)	$E\left[(p_f - E[p_f])^3 (p_m - E[p_m])^3\right]$	0.010 (0.002)	0.009 (0.002)
$E\left[(p_f - E[p_f])^3 (p_m - E[p_m])\right]$	0.015 (0.003)	0.015 (0.002)	$E\left[(p_f - E[p_f])^4 (p_m - E[p_m])^3\right]$	0.008 (0.002)	0.008 (0.001)
$E\left[(p_f - E[p_f])(p_m - E[p_m])^3\right]$	0.016 (0.003)	0.017 (0.002)	$E\left[(p_f - E[p_f])^3 (p_m - E[p_m])^4\right]$	0.008 (0.002)	0.008 (0.002)
$E\left[(p_f - E[p_f])^2 (p_m - E[p_m])^2\right]$	0.016 (0.003)	0.016 (0.002)	$E\left[(p_f - E[p_f])^4 (p_m - E[p_m])^4\right]$	0.007 (0.002)	0.001 (0.001)
		J-statistic:	2.7		
		P-value:	0.891		
		Sample size:	799		

Notes: This table reports generalized method of moments (GMM) estimates of moments of the joint distribution of job-specific white and black callback rates in the Arceo-Gomez and Campos-Vasques (2014) data. Estimates in columns (2) and (4) come from a shape-constrained GMM procedure imposing that the moments are consistent with a well-defined probability distribution. The J-statistic is the minimized shape-constrained GMM criterion function. The p-value come from a bootstrap test of the hypothesis that the model restrictions are satisfied.

Table A.IV: Sensitivity of moments and bounds to discretization grid

		Grid spacing	*J*-statistic (1)	Share discriminating (2)	Share disc., one call (3)	Share disc., two calls (4)	Share disc., three calls (5)
				A. Nunley et al. data			
$K_1 = 50$	$K_2 = 200$	Quadratic	23.13	0.266	0.483	0.549	0.678
$K_1 = 100$	$K_2 = 200$	Quadratic	23.10	0.366	0.692	0.708	0.841
	$K_2 = 400$	Quadratic		0.329	0.611	0.643	0.783
	$K_2 = 600$	Quadratic		0.319	0.588	0.627	0.765
$K_1 = 150$	$K_2 = 300$	Quadratic	23.09	0.401	0.765	0.770	0.880
	$K_2 = 600$	Quadratic		0.368	0.692	0.708	0.835
	$K_2 = 900$	**Quadratic**		**0.358**	**0.672**	**0.691**	**0.821**
		Rectangular		0.410	0.780	0.785	0.887
				B. Arceo-Gomez & Campos-Vasquez data			
$K_1 = 50$	$K_2 = 200$	Quadratic	3.5	0.220	0.762	0.709	0.570
$K_1 = 100$	$K_2 = 200$	Quadratic	2.8	0.218	0.738	0.717	0.599
	$K_2 = 400$	Quadratic		0.209	0.732	0.710	0.583
	$K_2 = 600$	Quadratic		0.209	0.730	0.708	0.579
$K_1 = 150$	$K_2 = 300$	Quadratic	2.7	0.220	0.727	0.718	0.606
	$K_2 = 600$	Quadratic		0.208	0.722	0.709	0.587
	$K_2 = 900$	**Quadratic**		**0.207**	**0.721**	**0.708**	**0.584**
		Rectangular		0.215	0.713	0.707	0.590

Notes: This table explores the sensitivity of our shape-constrained generalized method of moments (SCGMM) and linear programming bounds results to the number of grid points used to approximate the joint distribution of callback probabilities. K1 refers to the number of mass points used in the quadratic programming SCGMM step, while K2 refers to the number of mass points used in the linear programming bounds step. Quadratic grid spacing refers to the scheme described in Appendix A, and rectangular spacing refers to a grid with equally spaced points. Column (1) shows the minimized SCGMM criterion function for each value of K1. Column (2) displays the lower bound on the fraction of discriminating jobs for each combination of K1 and K2. Columns (3)-(5) show corresponding bounds conditional on the total number of callbacks. Panel A displays results for an application design with two white and two black applicants in the Nunley et al. (2015) data, and panel B displays results for the Arceo-Gomez and Campos-Vasquez (2014) data. Bold lines indicate the preferred specification used in the main text.

E-lit-010: From Callbacks to Careers: Do Audit Studies Predict Workforce Diversity?

Donato Onorato, Kamelia Stavreva, Adina Sterling and Emmanuel Yimfor | Columbia Business School Research Paper No. 6022958 (SSRN) | 2025-09

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6022958

About this source

The most direct test yet of the objection that a callback is not a job. The authors match the firms in Kline, Rose and Walters's large-employer correspondence experiment to LinkedIn employment records covering roughly eight million employees, and ask whether a firm's measured callback gap predicts who actually works there. It does: firms with larger callback gaps employ fewer Black workers, the implied employment gap is more than three times what the callback gap alone would predict, and the relationship is strongest among the firms that discriminated most in the experiment. Two readings are live and the paper cannot fully separate them - discrimination continuing past the resume screen, or Black applicants sorting away from firms with a reputation for it - but both are damaging to the claim that callback evidence is an artifact confined to the screening stage. It also speaks to the name critique in Section 3, since only about three percent of the Black employees identified carry the distinctive names used in the experiment and the gap still predicts outcomes for the broader workforce. It is a working paper and should be weighted accordingly, but it is the paper the external-validity argument now has to answer, and it supplies the second half of this section's hypothesis: the callback gap understates rather than overstates what happens in employment. NOTE: the full text could not be downloaded - SSRN blocks automated access - so nothing in this briefing may quote a number from this paper beyond what is stated in this description, which comes from the research pass rather than from the document.

Bound as a placeholder — the document itself could not be obtained. Retrieved 2026-09-10. sha256 6145f3553169800b...

Bound into an Economic Briefing · by Claude, may contain mistakes

Full content unavailable - SSRN (Elsevier Content Protection) returns HTTP 403 to every scripted route and serves a 'Content Blocked' page to a headless-browser print; the paper exists only on SSRN and no repository or author-page copy could be found.. Only partial content is included.

Full document begins on the next page.

S4-2: From Callbacks to Careers: Do Audit Studies Predict Workforce Diversity?

Donato Onorato, Kamelia Stavreva, Adina Sterling and Emmanuel Yimfor | Columbia Business School Research Paper No. 6022958 (SSRN) | 2025-09

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6022958

Full content unavailable - SSRN (Elsevier Content Protection) returns HTTP 403 to every scripted route and serves a 'Content Blocked' page to a headless-browser print; the paper exists only on SSRN - Unpaywall resolves it to the same blocked doi.org/SSRN link, Semantic Scholar has no record of the DOI, and neither Yimfor's nor Sterling's personal page hosts a loose PDF. A human signed in to SSRN can download it in one click.. Only partial content is included.

Source: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6022958

E-lit-011: Ban the Box, Criminal Records, and Racial Discrimination: A Field Experiment

Amanda Y. Agan and Sonja B. Starr | Quarterly Journal of Economics 133(1): 191-235 | 2018-02

library:Labor/Discrimination/agan starr_2018_btb field experiment_qje.pdf

About this source

Three findings bear on the section's hypothesis. First, the pooled callback gap describes no employer in particular: in the pre-BTB cross-section the black-white gap was 0.3 percentage points at employers who asked about criminal records (white applicants 2% more callbacks) and 3.1 points at employers who did not (33% more callbacks) — about ten times as large — and the race coefficient is statistically insignificant in the box sample. The same pooling problem appears geographically: the white advantage is 4.5 percentage points (37%) in New Jersey against 0.7 points (8%) in New York City. Second, the discrimination is concentrated in and driven by employers operating without individualized information: when box-remover firms lost the question, the race gap at those same firms jumped from 0.8 to 4.4 percentage points (7% to 43% in proportional terms), a difference-in-differences coefficient of -3.6 points, with a triple-difference estimate of 3.9 points once trends at unaffected employers are differenced out, and no comparable change at employers whose applications never changed. Third, on the translation from callbacks to labor markets, the authors run the arithmetic rather than leaving it to the reader: using the largest real-world race gap in felony convictions they could find (Bucknor and Barber), the callback pattern implies black callback rates fall 2.7 points and white rates rise 0.1 point, a 2.8-point widening of the gap that is a quarter of the overall 11.7% callback rate and would quintuple the pre-BTB gap; using NLSY97 conviction rates instead gives 4.2 points. They add that because the losses fall on black applicants without records, and most black men have no felony conviction, the experiment — where half the black applicants had one — understates the real-world harm. The mechanism is not accurate statistical inference: backing employer priors out of the callback rates implies employers act as if 87% of black and 16% of white applicants have felony convictions, a 71-point gap against a real-world gap of at most about 32 points, and employers do not respond at all to a GED, which predicts a record far better than race does. Doleac and Hansen's CPS results, cited approvingly, show the callback effect showing up as reduced employment of young black men.

Bound as the original PDF. Retrieved 2026-09-10. sha256 31058f8c55f6b501…

Bound into an Economic Briefing · by Claude, may contain mistakes

Full document begins on the next page.

E-lit-011: library:Labor/Discrimination/agan starr_2018_btb field experiment_qje.pdf

BAN THE BOX, CRIMINAL RECORDS, AND RACIAL DISCRIMINATION: A FIELD EXPERIMENT*

AMANDA AGAN AND SONJA STARR

"Ban the Box" (BTB) policies restrict employers from asking about applicants' criminal histories on job applications and are often presented as a means of reducing unemployment among black men, who disproportionately have criminal records. However, withholding information about criminal records could risk encouraging racial discrimination: employers may make assumptions about criminality based on the applicant's race. To investigate BTB's effects, we sent approximately 15,000 online job applications on behalf of fictitious young, male applicants to employers in New Jersey and New York City before and after the adoption of BTB policies. These applications varied whether the applicant had a distinctly black or distinctly white name and the felony conviction status of the applicant. We confirm that criminal records are a major barrier to employment: employers that asked about criminal records were 63% more likely to call applicants with no record. However, our results support the concern that BTB policies encourage racial discrimination: the black-white gap in callbacks grew dramatically at companies that removed the box after the policy went into effect. Before BTB, white applicants to employers with the box received 7% more callbacks than similar black applicants, but BTB increased this gap to 43%. We believe that the best interpretation of these results is that employers are relying on exaggerated impressions of real-world racial differences in felony conviction rates. *JEL Codes:* J15, J78, K31, K42.

*The authors gratefully acknowledge generous funding from the Princeton University Industrial Relations Section, the University of Michigan Empirical Legal Studies Center, and the University of Michigan Office of Research, without which this study could not have taken place. We thank Will Dobbie, Henry Farber, Alan Krueger, Steven Levitt, Alex Mas, Ezra Oberfield, Emily Owens, Alex Tabarrok, David Weisbach, Crystal Yang, and seminar participants at Princeton University, Rutgers University, the University of Chicago, the University of Michigan, Yale University, UCLA, the University of Pennsylvania, the University of Toronto, the University of Virginia, the University of Notre Dame, Northwestern University, Harvard, MIT, Harvard Law School, the Society of Labor Economists Annual Meeting, the IRP Summer Workshop, the American Law and Economics Association Annual Meeting, the NBER Summer Institute Labor Studies/Crime Session, and especially Larry Katz, Andrei Shleifer, and our anonymous referees for helpful comments. Finally, we thank every member of our large team of research assistants for their hard work and care, especially head RAs Louisa Eberle, Reid Murdoch, Emma Ward, and Drew Pappas, and our ArcGIS experts Linfeng Li and Grady Bridges. This experiment was initially registered with the AEA RCT Registry on April 16, 2015.

© The Author(s) 2017. Published by Oxford University Press on behalf of the President and Fellows of Harvard College. All rights reserved. For Permissions, please email: journals.permissions@oup.com
The Quarterly Journal of Economics (2018), 191–235. doi:10.1093/qje/qjx028.
Advance Access publication on August 2, 2017.

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

I. Introduction

Tens of millions of Americans—disproportionately including black men—have criminal records, and face resulting barriers to employment access. In an effort to help overcome those barriers, and thereby reduce racial disparities in employment, more than 150 jurisdictions and 25 states have recently passed "Ban the Box" (BTB) laws and policies (Rodriguez and Avery 2016). The "box" referred to in BTB (and hereinafter in this article) is the question on a job application form asking whether the applicant has been convicted of a crime, which is often accompanied by yes and no checkboxes. BTB prohibits employers from asking such questions on initial job applications or in interviews. Most BTB laws apply to public employers only, but nine states and several cities have now extended these restrictions to private employers.

BTB seeks to increase employment of people with criminal records. It is often further presented as an important tool for reducing race gaps in employment, especially for improving hiring of black men (Clarke 2012; Community Catalyst 2013; Pinard 2014; and Southern Coalition for Social Justice 2017), who have recently faced unemployment rates approximately double the national average (U.S. Bureau of Labor Statistics 2016).[1] But there is a plausible countervailing concern: absent individualized information, employers might instead rely on race-based assumptions about who is likely to have criminal records. If so, BTB could harm black men, in particular those with no records, who lose the ability to convey that fact on job applications.

We investigate BTB's effects via a field experiment. We submitted nearly 15,000 fictitious online job applications on behalf of young males to entry-level positions, before and after the effective dates of private-sector BTB laws in New Jersey (March 1, 2015) and New York City (October 27, 2015). We sent these applications in pairs matched on race (black and white) and also randomly

1. See, for example, Minnesota Department of Human Rights (2016): "The Ban the Box law can mitigate disparate impact based on race and national origin in the job applicant pool." New York City's public-sector BTB law was passed in 2011 as part of an initiative to improve employment of young black and Latino men (Astone, Katz, and Gelatt 2014). Civil rights organizations are also major BTB backers (Color of Change 2015; NAACP 2017).

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

varied whether the applicant had a felony conviction.[2] This design allows us to test, among other things, how employer reaction to race changes after BTB's adoption.

The basic premise of BTB laws—that criminal records are a major barrier to employment—finds support in prior research (Pager 2003; Holzer, Raphael, and Stoll 2006, 2007). BTB laws only delay (rather than permanently bar) employer access to criminal records; employers may still conduct criminal background checks near the end of the hiring process.[3] But the theory is that after meeting applicants in person, employers are less likely to treat records as being categorically disqualifying: "Rejection is harder once a personal relationship has been formed" (Love 2011). Thus, BTB seeks to help candidates with records to get their feet in the door. If BTB does increase hiring of people with records, that should help mitigate racial disparities in employment, because black men are more likely to have records (Shannon et al. 2011; Brame et al. 2014).[4]

On the other hand, BTB could also inadvertently encourage employers who lack criminal record information to rely on race as a proxy. Theories of statistical discrimination have long

2. We use "criminal record" and "felony conviction" (the type of record we varied experimentally) interchangeably here. Job application questions about records are overwhelmingly limited to convictions (not arrests), and usually to felonies.

3. In New Jersey and New York City, the two jurisdictions on which this study focuses, employers may conduct background checks any time after the first interview (New Jersey) or after a conditional job offer (New York). Nationally, a 2012 survey found that 69% of all employers conducted background checks (Society for Human Resources Management 2012), while a survey of 96 major retail chains in 2011 found that 97% of them performed background checks (Allen 2011). Some BTB laws also substantively restrict the role that criminal records can play in employers' ultimate decisions, but New Jersey's and New York's do not. New Jersey's law affects only the "initial employment application process" (N.J. P.L. 2014, Ch. 32). New York requires employers to consider whether a conviction is job-relevant, but this is a longstanding restriction that was unchanged by BTB (N.Y. Correction Law Sec. 752). All U.S. employers have also long been subject to similar restrictions (requiring nuanced assessments of criminal records) at the federal level, pursuant to the Equal Employment Opportunity Commission's interpretation of the Civil Rights Act of 1964 (EEOC 2012).

4. Brame et al. (2014) find that by age 23, 49% of black men have experienced an arrest, versus 38% of white men; Shannon et al. (2011) estimate that 25% of the U.S. black population has a felony conviction, compared with only 6% of the nonblack population. This disparate impact is why EEOC interprets race discrimination law to constrain employers' treatment of criminal records (EEOC 2012).

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

suggested that decision makers deprived of individualized information might rationally rely on group generalizations instead (Phelps 1972; Arrow 1973; Aigner and Cain 1977; Stoll 2009; Fang and Moro 2011). Alternatively, employers might rely not on accurate information but on exaggerated assumptions or stereotypes about group differences (e.g., Bordalo et al. 2016 provide a theory of the decision process producing stereotyping). Observational studies have examined whether racial disparities in employment are affected by use of employer background checks and expansion of Internet records databases, producing somewhat mixed results.[5] Other researchers have found evidence of increased reliance on race in reaction to other policies, such as drug testing or credit checks (Wozniak 2015; Bartik and Nelson 2016; Clifford and Shoag 2016).[6]

Our experimental design allows us to explore several related questions. First, we investigate whether employer callback rates vary by race and by felony conviction status. Second, we estimate how taking criminal history questions off the job application pursuant to BTB changes the race gap. This analysis exploits both the variation over time introduced by BTB and the fact that many employers, even before BTB, chose not to ask such questions on applications. We estimate BTB's effects on racial discrimination at affected employers, and we conduct triple-differences analyses that further difference out changes over the same period among similar employers whose applications were unchanged by BTB. In addition, we also test the effects of two other applicant characteristics that could potentially signal a criminal record: GED (versus an ordinary high-school diploma), and a one-year gap in employment history. Finally, we consider whether our results are consistent with rational statistical discrimination or whether employers are relying on exaggerated assumptions of criminality by race.

We report several key findings. The first supports BTB's core premise: when employers ask about them, felony convictions are

5. Bushway (2004) finds that Internet-based criminal records databases are associated with reduced race gaps in employment; in contrast, Finlay (2014) finds that while young black men without records benefit, these databases' net effect on young black men appears to be negative. Holzer, Raphael, and Stoll (2006) and Stoll (2009) find that surveyed employers who report that they use criminal records checks are more likely to hire African Americans.

6. Autor and Scarborough (2008) find that race gaps in retail hiring were unchanged by adoption of a test on which black applicants scored lower, suggesting that employers statistically discriminated before they used the test.

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

a major employment barrier. Applicants without convictions were 63% more likely to be called back than those with convictions (5.2 percentage points over a baseline of 8.2%). Second, however, BTB does appear to increase racial discrimination. The black-white gap in callbacks at BTB-affected employers grew by nearly 4 percentage points—a large expansion relative to our overall callback rate (11.7%). In our main specifications, this represented a sixfold increase in racial disparity: before BTB, white applicants received 7% more callbacks than similar black applicants, but after BTB this gap grew to 43% (or 45%, when trends at unaffected employers are further differenced out). The GED and employment gap variables, in contrast, did not significantly affect callback rates, and this did not change significantly after BTB.

The post-BTB increase in racial inequality in callback rates appears to come from a combination of losses to black applicants and gains to white applicants. In particular, black applicants without criminal records saw a substantial drop in callback rates after BTB, which their white counterparts did not see. Meanwhile, white applicants with criminal records saw a substantial increase in callbacks, which their black counterparts did not see. This pattern suggests that when employers lack individualized information, they tend to generalize that black applicants, but not white applicants, are likely to have records. Moreover, this phenomenon may contribute substantially to overall racial discrimination patterns. While we find an overall effect of race that is roughly in line with prior auditing studies, that effect is nearly absent at companies with the box before BTB, suggesting that a large share of observed racial discrimination may be driven by criminal record–based assumptions.

Disparate treatment based on race violates employment discrimination laws, whether it is based on accurate statistical generalizations or on inaccurate stereotypes.[7] However, the difference between the two mechanisms is important to understanding discrimination processes, which may in turn inform policy responses. We therefore seek to disentangle the two by applying our results to

7. Title VII of the Civil Rights Act of 1964, which prohibits race and sex discrimination in employment, does not permit otherwise illegal treatment based on group generalizations, even if they are empirically supported. For example, in *City of Los Angeles Dep't of Water and Power v. Manhart*, 435 U.S. 702 (1978), the Supreme Court held that an employer could not rely, in designing pension benefits, on the actuarial prediction that women live longer.

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

a simple model of employer decision making. This exercise implies that employers are relying on greatly exaggerated priors about racial differences in felony conviction rates, relative to the racial differences observed in a range of real-world studies. In addition, employers (even after BTB) do not give significant weight to applicant GED status even though real-world data indicate that having a GED instead of a regular high school diploma is a much stronger predictor of criminal records than race is. Together, these findings strongly imply that employers are not engaging in "rational" statistical discrimination, but rather are relying on stereotypes (see, e.g., Bordalo et al. 2016).

This study makes several distinct contributions to the literature. First, this is the first experimental study of BTB's effects, and indeed there is little empirical research of any sort on BTB. Two recent studies investigate BTB's effects on racial disparities using observational data from employment surveys. Doleac and Hansen (2016) find that in the Current Population Survey there is a decrease in employment for young, low-skill black and Hispanic men after BTB goes into effect. Shoag and Veuger (2016) use American Community Survey data and find in contrast that employment of black men increased after BTB, but do not break this down by age and/or education. These papers, unlike ours, focus primarily on public-sector BTB laws. We hope that our experimental method will shed important light on BTB's effects and inform ongoing legislative debates about BTB throughout the country.

Second, we use field experimental methods to contribute to the literature on statistical discrimination and stereotyping in employment, which has not generally used such methods.[8] Although our study is not a pure experiment (a key variable, whether the application asks about records, is not manipulated), our ability to perfectly observe and randomize all of our fictional applicants' characteristics allows us to avoid many of the most likely threats to causal inference that affect purely observational research.

Finally, we make a methodological contribution to the literature on auditing, which has for decades been a central tool for empirical research on discrimination in employment, housing, lending, and other areas. In the employment context, auditing involves

8. See List (2004) for an experimental approach to statistical discrimination in another context, sports card trading.

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

varying characteristics of interest about a job candidate while holding other characteristics constant. It has been used to test employment discrimination based on race, gender, length of unemployment spell, age, commute time, and type of postsecondary education (Neumark et al. 1996; Riach and Rich 2002; Bertrand and Mullainathan 2004; Lahey 2008; Oreopoulos 2011; Neumark 2012; Kroft, Lange, and Notowidigdo 2013; Farber, Silverman, and Von Wachter 2015; Neumark, Burn, and Button 2015; Deming et al. 2016), as well as the effect of criminal records (Pager 2003; Pager, Bonikowski, and Western 2009; Baert and Verhofstadt 2015; Uggen et al. 2014; Decker et al. 2015). But auditing has previously only been used to obtain a one-time snapshot of discrimination patterns—to our knowledge, ours is the first study to use it to assess the effects of a policy. Because researchers cannot randomize the application of the policy itself, using auditing to assess policies requires combining the field-experimental approach with additional methods of causal inference—in this case, difference-in-differences analysis. We believe that combining auditing with quasi-experimental analysis of policy changes enriches the study of discrimination.

II. Experimental Design

We submitted online job applications on behalf of fictitious job applicants to low-skill, entry-level job openings both before and after BTB went into effect in New Jersey and New York City. New Jersey's version of BTB, the Opportunity to Compete Act, became effective March 1, 2015. We submitted applications in New Jersey between January 31 and February 28, 2015 (the pre-BTB period) and between May 4 and June 12, 2015 (the post-BTB period). New York City's BTB law went into effect on October 27, 2015. We submitted applications in New York City between June 10 and August 30, 2015 (the pre-BTB period) and between November 30, 2015, and March 31, 2016 (the post-BTB period). Our main outcome of interest—the "callback"—is whether an employer left a voicemail or email requesting that the applicant contact them or requesting an interview. These phone calls and emails were tracked for eight weeks from the application date.[9]

9. In NJ, pre-BTB period data collection finished after BTB went into effect, but the applications were submitted before it went into effect, so the applications to employers with the box did contain criminal record information.

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

II.A. Choosing Employers and Job Postings

Our subjects were private, for-profit employers. We relied on two main sources to locate openings. First, we searched two major online job boards: indeed.com, the largest U.S. job website, and snagajob.com, the largest site focused on hourly employment.[10] Second, we also directly searched the employment websites of chain businesses meeting certain criteria.[11] We searched for jobs requiring little work experience, no postsecondary education, and no specialized skills: predominantly crew-member restaurant and retail jobs. We focus on these sectors, and specifically on chains, because they typically require online job applications, rather than just evaluating résumés (which do not have a "box" that can be banned). These sectors also are likely to attract applicants with criminal records, who disproportionately have limited work experience and education. The applications were filled out with the help of a large team of University of Michigan student research assistants (RAs). While submitting job applications, the RAs filled out a spreadsheet that indicated details about the employer, position applied to, and the questions asked on the application.

In the post-BTB period, most applications were sent to employers that we had already applied to in the pre-BTB period; these were supplemented with some additional stores from the same chains. Stores thus received up to four applications total, one pair in each period. It was sometimes impossible to send a

10. Prior auditing studies have often identified jobs based on newspaper classified ads; today, these have largely been displaced by online sites, or are included in multisite aggregators like Indeed (Del Castillo 2016).

11. In New Jersey, we applied to chains with at least 30 locations and 300 employees in the state. In New York, we applied to these same chains, plus other chains with at least 20 locations in the city. We excluded employers that did not use online job applications, although the vast majority of chains meeting those size criteria (or advertising on Snagajob or Indeed) do use them. We excluded a few chains due to extremely arduous online application processes, and a few that targeted an overwhelmingly female clientele. Finally, some employers required full SSNs on job applications. For ethical reasons, we avoided using potentially real SSNs, instead using invalid SSNs beginning with 9xx or 666. Some employers had systems that automatically detected these invalid SSNs, and we excluded those businesses. If setting up such a system were correlated with special interest in criminal records, then excluding this pool could reduce our estimates of the effects of a criminal record. However, among employers we did apply to, there was no correlation between whether employers asked for an SSN at all and whether they asked about criminal records.

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

complete set of four applications, usually because the store was hiring in one period but not the other. In addition, a few RA assignments were not completed before BTB's effective date.[12] As a result, the sample composition by chain, jurisdiction, and specific store is not identical across periods. We address this concern below.

II.B. Applicant Profiles

Our fictitious applicants are all male and approximately 21 to 22 years old.[13] The RAs filled out applications based on profiles that we created using the Résumé Randomizer program created by Lahey and Beasley (2009). Each applicant profile included a name, phone number, address, employment history consisting of two prior jobs, unique email address, two references with phone numbers, information on high school or GED programs, felony conviction status and information about the criminal charge, formatted résumé, and answers to many other routine application questions, such as job availability and pay sought (minimum wage).[14]

The profiles were created in pairs of one black and one white applicant, which were assigned to the same store in the same time period. There was a time lag within pairs, with order randomized. In addition to race, other treatment dimensions that we randomized were:

(i) Has felony criminal conviction or not
   a. (Conditional on conviction): convicted of property crime or drug crime
(ii) Has one-year employment gap versus a zero- to two-month gap (referred to as "no gap" below) between the two past jobs
(iii) GED or high school diploma

12. This occurred mainly in the New Jersey pre-BTB period, which had to be completed relatively quickly. In New Jersey, we filled in these gaps in the post-BTB period whenever possible. In New York City, our pre-BTB wave was quite comprehensive, so we limited the post-BTB wave to stores that we had sent at least one application to pre-BTB.

13. Employers rarely ask about age or high school graduation year due to age discrimination laws but could potentially infer it via length of work history.

14. It was not possible for the profiles to anticipate every question asked, so we relied on the RAs' judgment but provided detailed training about what employers are generally looking for.

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

Race is indicated via applicant names, as discussed in Section II.C. The felonies we gave our applicants were nonviolent and fairly minor—either property crimes (e.g., shoplifting or receiving stolen property) or drug crimes (e.g., controlled substances possession). Like race, the employment gap and GED variables potentially could be seen by employers as proxies for criminal history.[15]

We chose 40 geographically distributed cities/towns in New Jersey and 44 neighborhoods throughout New York City's boroughs to serve as "centers" where the applicants' addresses would be located; each center then served as a base for application to nearby employers.[16] All applicant addresses were in racially diverse, lower- to middle-class neighborhoods. Other applicant characteristics such as work history, address within center, and high school were designed to have similar connotations, but randomized among similar options so they did not appear identical. Most applicant profiles (59% of the sample) were sent to only one store, but we sometimes used the same pairs to apply to multiple nearby locations of the same chain, as real-world applicants might do.[17]

15. As of 2005, 13.6% of GEDs were issued in state and federal prisons (Heckman and LaFontaine 2010). The relationship between GED, race, and criminal records is further addressed in the discussion. The one-year employment gap is meant to signal potential time spent incarcerated. Absence of a gap does not necessarily imply no conviction, however, because offenders are often not incarcerated. Among all individuals charged with felonies in state courts, 62% are not detained before trial; 27% of those convicted receive no incarceration, and of those sentenced to incarceration, approximately 24% received jail sentences of one to three months (Reaves 2013). Incarceration rates are presumably lower yet among first-time offenders with minor felonies, like our fictional applicants.

16. In New Jersey, we assigned each municipality in the state to its nearest center, minimizing distance. In New York City, because distances are much smaller generally, we prioritized distributing the locations of each chain across centers, and minimized distance within equal-distribution constraints.

17. Our criteria for this grouping differed between New Jersey and New York City. In New Jersey, we were concerned that the same hiring managers might cover multiple locations of chains and might become suspicious upon noticing groups of similar applicants coming within a short time from the same nearby town. Accordingly, we used the same applicant profiles for all locations that were assigned to a given center, as though just one applicant was applying. In New York, our concerns were different: the centers are not towns and likely appear less distinctive to managers, and we had more available time before BTB's effective date, so we were able to space out the timing of our applications (generally by a month or more). Thus, in New York we chose to increase power by sending each application to only one location, except for the largest five chains (in which we sent each applications to up to two or three stores).

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

For more details on applicant profiles and application procedures, see Online Appendix A1.[18]

II.C. Indicating Applicant Race

Race is the central characteristic of interest in our study, and we signal race by the name of the applicant (Bertrand and Mullainathan 2004; Oreopoulos 2011). To identify racially distinctive names, we used birth certificate data for babies born between 1989 and 1997 from the New Jersey Department of Health (NJDOH), which encompasses the cohort that would include our applicants (Center for Health Statistics n.d.). We then chose first and last names that were racially distinctive (meeting threshold requirements for the percentage of babies given that name who were black or non-Hispanic white) and common (meeting threshold frequency requirements).[19] Applicants were assigned random first and last names from the appropriate lists, which are provided in Online Appendix A2. The combination of distinctive first and last names should produce a very strong racial signal: according to the birth certificate data, 97% of persons with first and last names on our "black" list are black, and 92% of persons with first and last names on our "white" list are white.

One concern is that racially distinctive names could also signal socioeconomic status (SES), which employers may believe to be correlated with productivity (Fryer and Levitt 2004). However, our applications provided a great deal of concrete SES information to employers, including work histories, education, current neighborhood, and high school location. With all this information available, employers likely would not need to rely on names to draw SES inferences. To further mitigate this concern, we chose common names (avoiding socioeconomic connotations associated with unusual names) and also limited our white name list to those

18. All appendix material can be found in the Online Appendix.

19. Because blacks are a much smaller fraction of the population, these thresholds varied by race: the minimum percentages were 80% for white first names, 85% for white last names, and 70% for black first and last names, while the minimum frequencies were 450 for white first names, 150 for white last names, 150 for black first names, and 100 for black last names. We eliminated a few first names that were not distinctively male or that had strong associations with a particular religion, to avoid confounding race's effects with other variables. A heavily overlapping list would have been chosen had we followed the approach of Bertrand and Mullainathan (2004) or Fryer and Levitt (2004).

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

below the white SES median, as measured by maternal education, the best available indicator.[20] Finally, racially distinctive names are very common and do not point to an individual being a high- or low-SES outlier within their race.[21] Thus, even if employers do make assumptions about SES based on such names, similar assumptions would affect a large fraction of real-world job applicants.

III. Summary Statistics and Main Effects of Applicant Characteristics on Employer Callbacks

We submitted 15,220 applications, of which 14,637 are in our analysis sample.[22] This includes 6,401 applications in New Jersey and 8,236 in New York City. The applications were sent to 4,291 establishments ("stores") in 293 chains. We begin with summary statistics and then analyze the main effects of applicant characteristics on employer callbacks. The summary statistics and results presented in the tables and figures below combine both jurisdictions; in Online Appendix A5 and A6, we replicate several of the tables and figures for New Jersey and New York separately.

20. It was not possible to create lists that were equivalent on SES; virtually every distinctively white name averages higher than virtually every distinctively black name, due to socioeconomic stratification by race. Although some SES gap remains, it is very similar to the overall SES gap between black and white citizens—that is, choosing distinctive names did not amplify the gap (even if employers were to rely on names to signify SES).

21. In our birth certificate sample, 47% of black children have a racially distinct first name and 36% have a racially distinct last name (as we define distinctiveness, see note 20), while 35% of white children have a racially distinct first name and 65% have a racially distinct last name.

22. The remaining 580 observations (3.8% of those we sent) were dropped for several reasons. First, when an entire chain was applied to only in one period, our key treatment variable (*Box Remover*) could not be coded. Second, some stores had inconsistencies within one or both rounds as to whether the box was present, generally either because of precompliance before BTB's effective date (occurring between the two applications) or because of RA mistakes (missing disclaimers saying not to answer the criminal record question). In these cases we discarded the observation that was an outlier from the overall chain norm (including the RA-mistake observations, or in the precompliance cases, the later, nonbox observation). Third, we also dropped some businesses (about 1% of the sample) that appeared, mysteriously but presumably due to an administrative mistake, to add the box after BTB and therefore could not be coded as 0 or 1 on the *Box Remover* variable. We add these back in in a robustness check below, with the coding of −1.

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

TABLE I

MEANS OF APPLICANT AND APPLICATION CHARACTERISTICS AND CALLBACK RATES BY PERIOD

	Pre-BTB	Post-BTB	Combined
Characteristics			
White	0.502	0.497	0.500
Conviction	0.497	0.513	0.505
GED	0.498	0.502	0.500
Employment gap	0.492	0.504	0.498
Application has box	0.366	0.036	0.199
Results			
Callback rate	0.109	0.125	0.117
Interview req.	0.060	0.067	0.063
Callback rate by characteristics			
Black	0.099	0.111	0.105
White	0.120	0.139	0.129
GED	0.106	0.127	0.117
HSD	0.113	0.122	0.118
Emp. gap	0.110	0.126	0.118
No emp. gap	0.109	0.124	0.116
N	7,245	7,392	14,637

Notes. "Callback" means application received a personalized positive response from the employer (either voicemail or e-mail). Interview request means the positive response message specifically mentioned an interview. "Application has box" means that the application asked about criminal records. "Employment gap" or "Emp. gap" means an 11–13-month employment gap in work history, while "No emp. gap" means a 0–2-month gap.

III.A. Summary

Initial summary statistics are presented in Table I, by period and overall. As expected, approximately 50% of our applications had each of our randomized characteristics, but our other characteristic of interest—whether the application had the criminal-record "box"—could not be randomized. Among pre-BTB applications, 36.2% had the box; after BTB, 3.6% still had it (noncompliers). Thus, 33% of the sample (4,793 observations, 1,383 employers, and 71 chains) consisted of "box removers": employers that had the box before BTB, but not after. The rate at which employers had the box before BTB may seem surprisingly low, given earlier studies finding rates as high as 80% (see Uggen et al. 2014, reporting results from 2007–2008). A plausible explanation for this difference is the recent success of the BTB movement, which has affected employers (especially national chains) even in jurisdictions without a BTB law. Most of the nonbox employers have no box on their application in any jurisdiction, indicating

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

TABLE II
CALLBACK RATES BY CRIME STATUS FOR STORES WITH THE BOX IN THE PRE-BTB PERIOD

	No crime	Crime	Property	Drug	Combined
Callback rate	0.136	0.085	0.084	0.085	0.110
Callback black	0.131	0.086	0.091	0.081	0.109
Callback white	0.140	0.083	0.077	0.089	0.111
N	1,319	1,336	703	633	2,655

Notes. Sample restricted to pre-BTB period applications where the application asked about criminal records. Callback implies application received a personalized positive response from the employer.

that very few reflect early compliance with the New Jersey or New York BTB laws specifically.

Overall, 11.7% (1,715) of our applications received callbacks.[23] This rate was higher in the post-BTB period (12.5% versus 10.9%), and lower in New York City than in New Jersey (9.4% versus 14.7%; see Online Appendix Tables A5.1a and A6.1a). Among the callbacks, about 55% specifically mentioned an interview (though many others were likely seeking interviews even if the message did not so specify). The race gap in callback rates grew from 2.1 percentage points in the preperiod to 2.8 percentage points in the postperiod; these averages do not differentiate box-remover employers from others and mask large changes occurring at box-remover employers, as shown below. Callback rates hardly differed by GED/diploma status or employment gap.

Table I does not break down callback rates by criminal record status because criminal record is unobserved by most employers (those without the box). Table II thus shows separate summary statistics limited to pre-BTB applications to employers with the box. Callback rates were 60% higher for applicants without criminal records (5.1 percentage points, over a base rate of 8.5%). Applicants with drug and property crime convictions had similar callback rates—perhaps surprisingly, as one might have expected employers to be particularly concerned about potential employee theft. When employers asked about records, we saw essentially no race gap in callback rates (11.1% for whites, 10.9% for blacks).

23. This rate is similar to other recent audit studies: Kroft, Lange, and Notowidigdo (2013) had a positive response rate of 11.6%; Farber, Silverman, and Von Wachter (2015) had a 10.4% callback rate; Deming et al. (2016) had an 8.2% callback rate.

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

TABLE III

EFFECTS OF APPLICANT CHARACTERISTICS ON CALLBACK RATES

	(1)	(2)	(3)
White	0.024***	−0.001	−0.001
	(0.006)	(0.009)	(0.009)
Conviction	−0.014**	−0.052***	
	(0.005)	(0.012)	
GED	−0.004	0.010	0.010
	(0.005)	(0.014)	(0.013)
Employment gap	0.002	0.011	0.011
	(0.005)	(0.010)	(0.010)
Pre-BTB period	−0.015		
	(0.010)		
Drug conviction			−0.050***
			(0.013)
Prop. conviction			−0.054***
			(0.014)
N	14,637	2,918	2,918
Sample	All	Box	Box
Chain FE	Yes	Yes	Yes
Center FE	Yes	Yes	Yes

Notes. Dependent variable is whether the application received a callback. Standard errors clustered on chain in parentheses. Chain and geographic center fixed effects are included in all regressions. White is an indicator for race (versus black), Conviction is an indicator for whether the applicant has a felony conviction, GED is an indicator for having a GED (versus a regular high-school diploma), and Employment gap is an indicator for whether the applicant has an 11–13-month gap in work history between the previous two jobs (versus a 0–2-month gap). "Drug conviction" and "Prop. conviction" break the Conviction variable into a categorical variable based on crime type (drug versus property crime); no conviction is the base category. The box sample is employers with the box on their application. *10%, **5%, and ***1% significance level.

The advantage for applicants without records was slightly larger for white applicants (5.7 percentage points, or 69% above the base rate of 8.3%) than for black applicants (4.5 percentage points, or 52% above the base rate of 8.6%), although regressions not shown here show that the race-record interaction is not statistically significant.

III.B. Regression Estimates of Main Effects of Applicant Characteristics

Table III provides regression estimates of the main effects of race, record, GED/diploma status, and employment gap on callback rates; the regressions also include fixed effects for the chain (with the smallest chains grouped by business category) and the

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

geographic center.[24] These estimates parallel the summary statistics, which is not surprising given that applicant characteristics were distributed randomly. Column (1) shows results for the full sample. White applicants were about 2.4 percentage points (23%) more likely to receive a callback. In contrast, callback rates did not vary based on the GED and employment gap variables.

To assess the effect of having a criminal record when employers observe it, we show analyses in columns (2) and (3) that are limited to employers with the box. Column (2) shows a 5.2-percentage-point criminal record effect ($p < .01$), which translates into a 63% higher callback rate for applicants without records compared to the 8.2% baseline for applicants with records. Column (3) shows that this effect was similar for property and drug crimes. In Agan and Starr (2017a), we report additional statistics detailing the criminal record effect, including variations by applicant race, industry, local demographics, and crime rates. Meanwhile, the main effect of race is economically and statistically insignificant in the box sample, a point further examined in the remainder of this article, which assesses the effect of the box on racial discrimination.

In Online Appendix A4 we show that the main effects of race and crime are robust to several alternative specifications and samples.[25] The biggest differences are geographic. The "white" effect is far larger in New Jersey (4.5 percentage points, or 37% more callbacks for whites) than in New York City (0.7 percentage points, or 8% more callbacks for whites).[26] The criminal record effect, in contrast, is larger in New York City, at least in proportional terms.[27] At box employers, applicants without records receive 45% more callbacks than those with records in New Jersey; in New York City, applicants without records receive 78% more callbacks.

24. All the results shown in Table III are for both periods combined (unlike Table II and Figure II, which were for the pre-BTB period only), but the regression results look similar if only the pre-BTB observations are used.

25. These robustness checks parallel those discussed below concerning the analyses of BTB's effects (Table VI).

26. In a separate article, we further explore the geographic variations in the race effect. We find that businesses in whiter neighborhoods much more strongly favor white applicants, suggesting that New York City's greater racial diversity could partially (but not fully) explain its smaller race gap. For a preliminary version, see Agan and Starr (2016).

27. New Jersey's callback rate was higher, so similar percentage-point gaps translate into different proportional effects.

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

IV. The Criminal Record Box and Racial Discrimination

In this section, we show difference-in-differences analyses that shed light on the effect of criminal record information on racial discrimination in callbacks. We exploit two different sources of variation in whether employers have the box. First, in Section IV.A, we briefly examine the cross-sectional variation between employers that (before BTB) chose to ask about criminal records ("box employers") and those that did not. Second, in Section IV.B, we assess the temporal change after BTB for employers that had the box before BTB and then removed it (and show that no similar change existed for employers unaffected by BTB). Third, in Section IV.C, we employ a triple-differences analysis that combines both these sources of variation.

IV.A. Cross-Sectional Difference-in-Differences Estimates

In Table III, column (1), we compare race gaps between employers that do and do not have the criminal record "box." We employ a simple difference-in-differences specification for the probability that applicant i to store j receives a callback:

$$Callback_{ij} = \alpha + \beta_1 Box_j + \beta_2 White_i + \beta_3 Box_j \; x \; White_i + \mathbf{\Gamma} \boldsymbol{X_i} + \epsilon_{ij}. \quad (1)$$

Box_j indicates whether store j has the box, $White_i$ indicates applicant race, $Box_j \; x \; White_i$ is the interaction of those variables, and $\boldsymbol{X_i}$ is a vector of control variables: GED, employment gap, and geographic center.

In column (1), we present results from equation (1) in the pre-BTB sample only. This sample includes 7,245 observations in 3,874 stores and 293 chains. The $Box_j \; x \; White_i$ coefficient in column (1) indicates that the black-white gap is 2.8 percentage points larger among nonbox employers ($p < .05$) in the preperiod. Among box employers, the race gap is just 0.3 percentage points (in proportional terms, white applicants received 2% more callbacks than black applicants did). Among employers without the box, the gap was about 10 times as large, 3.1 percentage points compared to a base rate for black applicants of 9.4%; in proportional terms, white applicants received 33% more callbacks than black applicants did.

Figure I provides a visual representation of this cross-sectional comparison that further breaks down box employers'

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

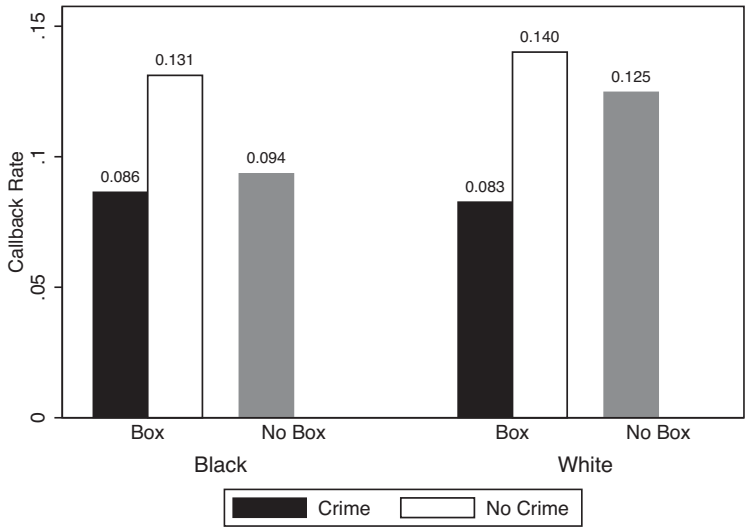


FIGURE I

Callback Rates by Race, Crime, and Box: Preperiod Applications Only

This figure compares callback rates within the preperiod before Ban the Box went into effect, comparing applications with the criminal record question box and those without the box.

callback rates by applicants' conviction status. This figure shows that at box employers, while conviction status itself dramatically affects callback rates, race makes little difference. Among applicants with records, black callback rates are slightly higher (8.6% versus 8.3%), while among applicants without records, white callback rates are higher (14.0% versus 13.1%); neither difference is significant. Again, only at nonbox employers does a significant race gap emerge (12.5% versus 9.4% for white and black applicants, respectively).

This analysis suggests that employers who lack individualized information might be relying on race-based assumptions about criminal record status. Conversely, the absence of a race effect among box employers is intriguing, providing a sharp contrast from other auditing studies, which do not differentiate by "box" status and which generally do find race gaps (as we did in the full sample; see Table III, column (1)). A possible implication is that a substantial share of the racial discrimination observed in other studies might be driven by employers who lack criminal record information and make negative assumptions about black applicants' criminality.

However, the two groups of employers could differ in other ways, such that the patterns we observe do not relate to the

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

box. In Online Appendix A3, we show that the two groups do not vary substantially across most observable characteristics: index crime rates of store neighborhood, black and white percentages of the neighborhood population, average number of employees, and average sales volume.[28] Retail employers are noticeably more likely to have the box than other employers (which are mainly restaurants); however, additional regression analyses (not shown here) indicate that this difference is not what drives the $Box_j \, x \, White_i$ interaction.[29] Overall callback rates were also nearly identical at the two groups of employers. Still, unobservable differences between the two groups are possible, so the cross-sectional analysis is only suggestive; more rigorous causal identification is left to the analyses below.

IV.B. Temporal Difference-in-Differences Estimates of BTB's Effects

Because BTB introduced exogenous variation in whether employers have the box, we need not rely only on comparing different groups of employers. In columns (2), (3), and (4) of Table IV, we present difference-in-differences analyses exploiting this temporal variation, limited to "box remover" employers: those that had the box before BTB and removed it afterward. Like the column (1) analysis, these regressions implement equation (1). However, the source of variation in the *Box* variable is now time (and the intervening policy change), rather than cross-sectional variation among businesses. Thus, Box_j becomes equivalent to an indicator for the pre-BTB period.

In column (2), the analysis is limited to box-remover stores to which we were able to send a complete set of four applications (the "box-remover balanced sample"): 3,712 observations in 928 stores and 62 chains. This limitation means that the sets of employers being compared are identical before and after BTB, so there is no cross-sectional variation being inadvertently introduced due, for example, to different openings being available in different

28. Because the industry difference may be correlated with other characteristics, we also run a simple regression of whether a store has the box on all these characteristics, showing that the only characteristic with a substantial and significant effect is the retail indicator. See Online Appendix A3.

29. Specifically, we add a $Retail_j$ indicator, interacted with $White_i$, to the column (1) regression. The $Box_j \, x \, White_i$ coefficient only increases in magnitude, to 3.4 percentage points ($p < .05$). Similar analyses also show that the small differences in the other observable employer characteristics in Online Appendix A3 do not explain the $Box_j \, x \, White_i$ effect.

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

TABLE IV

EFFECTS OF THE BOX ON RACIAL DISCRIMINATION: DIFFERENCE-IN-DIFFERENCES

	(1)	(2)	(3)	(4)	(5)
Box × white	−0.030**	−0.036**	−0.033**	−0.027**	0.002
(*White* × *pre, column* (5))	(0.015)	(0.014)	(0.014)	(0.013)	(0.014)
White	0.032***	0.044***	0.040***	0.123	0.022**
	(0.012)	(0.013)	(0.012)	(0.132)	(0.009)
Box	0.015	0.003	−0.002	−0.345**	−0.016
(*Pre, column* (5))	(0.024)	(0.015)	(0.013)	(0.139)	(0.017)
N	7,245	3,712	4,794	4,794	7,476
Controls	Yes	Yes	Yes	Yes	Yes
Center FE	Yes	No	Yes	Yes	No
Chain FE	No	No	No	Yes	No
Post × chain FE	No	No	No	Yes	No
White × chain FE	No	No	No	Yes	No
Box variation	Cross-section	Temporal	Temporal	Temporal	None
Sample	Pre-BTB	Box remover -balanced	Box remover -full	Box remover -full	Other empl. balanced

Notes. Standard errors clustered on chain in parentheses. Dependent variable is whether the application received a callback. Box removers are stores that had the box in the pre-BTB period and removed it after BTB. "Box removers-balanced" consists of box remover stores to which we sent exactly four applications, one white/black pair in each period. Fixed effects can include geographic center, chain, post × chain, and white × chain, and are included as indicated; note that because of the inclusion of interacted fixed effects in column (4), the white and box coefficients are not meaningful. Controls are whether the applicant had a GED (versus regular high-school diploma) and whether he had an employment gap. Box variation indicates the source of variation in the box variable: "Cross-section" means the variation comes from a comparison of box and nonbox stores in the preperiod; "Temporal" means the variation is pre- and post-BTB, triggered by the implementation of the BTB policy. In the last column, which is shown as a comparison point, there is no box variation; the pattern over the same time period is shown for companies that did not change their job applications. *10%, **5%, and ***1% significance level.

seasons. Accordingly, the vector of controls in column (2) includes GED status and employment gap but not the geographic center, which is already perfectly balanced. In column (3), we show nearly the same analysis but in the full sample of box removers, adding back the center fixed effects. This sample is larger (4,794 observations in 1,383 stores and 71 chains), at the cost of some imbalance in the employers represented across time periods. In column (4), also carried out in the full box-remover sample, we add chain fixed effects and interact them with Box_j and $White_i$, which accounts for the pre/post imbalances in chains but not individual stores.[30]

Each of these analyses shows that racial discrimination increased substantially when these companies removed the box to

30. In this regression, the main effects of *Box* and *White* do not have a meaningful interpretation because those variables' effects are spread across the interacted fixed effects. The principal term of interest, *Box* × *White*, retains its interpretation.

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

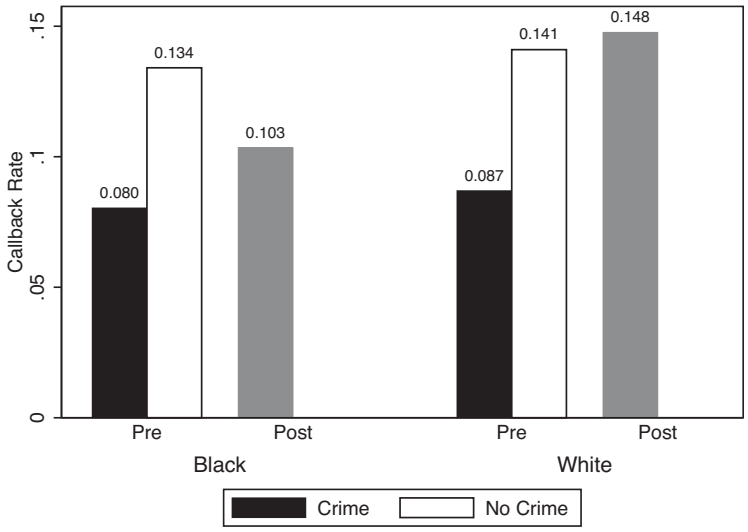


FIGURE II

Callback Rates by Race, Criminal Record, and Period: Balanced Box Removers Only

This figure compares callback rates before and after Ban the Box went into effect, among companies that had the criminal record question box before BTB and removed it afterward, in the balanced sample only (i.e., stores to which we sent complete application pairs in both the pre-BTB and post-BTB periods).

comply with BTB. The $Box_j \times White_i$ coefficient is −3.6 percentage points in the balanced sample (column (2))—that is, when these employers had the box, the race gap was 3.6 percentage points smaller than after it was removed. In the pre-BTB period, the white callback rate was about 0.8 percentage points higher than the black baseline of 10.7%; in proportional terms, whites received 7% more callbacks. In the post-BTB period, after these companies dropped the box, this race gap ballooned. The white callback rate was 4.4 percentage points higher than the black baseline of 10.4%; in proportional terms, whites now received 43% more callbacks. The $Box_j \times White_i$ coefficient is slightly smaller in the full sample (3.3 percentage points and 2.8 percentage points without and with the interacted chain fixed effects, respectively), but the overall pattern is very similar: a multifold enlargement of the black/white gap after BTB. The effects are all statistically significant, with p-values ranging from .01 to .03.

Figure II represents these patterns visually (for the balanced sample of box-removers).[31] Like Figure I, it further decomposes

31. The figure looks very similar if the full box-remover sample is used instead.

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

the callback rate when employers have the box (here labeled as "Pre") based on criminal record status. The pattern is similar to that observed in Figure I. Before BTB, callback rates are only slightly higher for white applicants (whether with or without records), but they become substantially higher (15% versus 10.3% for black applicants) after BTB.

The temporal difference-in-differences analysis is more causally rigorous than the cross-sectional comparison: it compares results across the same employers, just a few months apart, facing identical pools of fictional applicants. It seems unlikely that non-box-related differences would explain such a sharp increase in the race gap, and this intuition is further supported by the striking similarity to the cross-sectional results. Still, the estimates could potentially be confounded by trends unrelated to BTB. For example, if adoption of BTB reflects a motivation to address racial disparities in employment, that motivation could in theory affect disparity trends in other ways.[32] An initial check of this possibility is to run a similar difference-in-differences analysis in the sample of employers whose applications were unchanged after BTB (predominantly employers that never had the box). These results are in column (5), which reflects the same specification as column (2), carried out in the balanced sample.[33] In sharp contrast to the box-remover employers, among other employers there was essentially no change in the black-white gap between the pre-BTB and post-BTB periods (indeed, the sign is flipped, though this difference is insignificant).[34] To further address these concerns, we turn to the triple-differences analysis.

IV.C. Triple-Differences Estimates of BTB's Effects

Here, we further analyze the causal effect of BTB on racial discrimination via a difference-in-difference-in-differences analysis, which exploits both sources of variation (cross-sectional and temporal) discussed above. This analysis compares the change

32. Seasonal variation is also possible, although this possibility is mitigated by the fact that the timing of the NYC and NJ studies was nearly seasonally opposite.

33. This analysis substitutes *pre* for *box* in equation (1), but it is still directly parallel to columns (2) through (4), which could have those variables labeled either way (within the box-remover sample, *pre* is equivalent to *box*).

34. Column (5) does show an overall higher callback rate in the post-BTB period, which is also seen in the box-remover stores, and which is presumably unrelated to BTB.

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

TABLE V

EFFECTS OF BAN THE BOX ON RACIAL DISCRIMINATION: TRIPLE DIFFERENCES

	(1)	(2)	(3)
Box Remover × post × white	0.039*	0.040**	0.035*
	(0.020)	(0.018)	(0.018)
Post × white	−0.002	−0.006	−0.006
	(0.014)	(0.012)	(0.013)
Box Remover × post	−0.019	−0.011	
	(0.023)	(0.019)	
Box Remover × white	−0.017	−0.021	
	(0.015)	(0.014)	
Box Remover	0.016	0.009	
	(0.028)	(0.024)	
White	0.024**	0.028**	0.098
	(0.012)	(0.011)	(0.129)
Post	0.016	0.012	0.339**
	(0.017)	(0.015)	(0.139)
N	11,188	14,637	14,637
Controls	Yes	Yes	Yes
Center FE	No	Yes	Yes
Chain FE	No	No	Yes
Post × chain FE	No	No	Yes
White × chain FE	No	No	Yes
Sample	Balanced	Full	Full

Notes. Standard errors in parentheses clustered on chain. Dependent variable is whether the application received a callback. A "box remover" store is one that had the box in the pre-BTB period and then removed it due to BTB. The balanced sample consists of stores to which we sent exactly four applications, one white/black pair in each period. Controls are whether the applicant had a GED (vs. regular high-school diploma) and whether he had an employment gap. Fixed effects for the chain, post × chain, white × chain, and geographic center are included as indicated. *10%, **5%, and ***1% significance level.

in racial discrimination after BTB at box-remover employers to changes over the same period at other employers. This analysis will "difference out" the effect of any non-BTB temporal differences so long as they affect both sets of employers similarly. Similarly, the analysis effectively controls for unobserved cross-sectional differences between the two groups of employers so long as they are time-invariant over the period in question. The change in the race gap that remains after this differencing-out is interpreted as the causal effect of a chain's compliance with BTB.

In Table V, columns (1) and (2), which are carried out in the balanced sample and the full sample, respectively, we apply the following triple differences estimating equation for the probability that applicant i to store j in time period t receives a

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

callback:

$$\begin{aligned} Callback_{ijt} = \alpha + \ & \beta_1 White_i + \beta_2 Post_t + \beta_3 Box Remover_j \\ & + \beta_4 White_i \ x \ Post_t + \beta_5 White_i x \ Box Remover_j \\ & + \beta_6 Post_t \ x \ Box Remover_j \\ & + \beta_7 Box Remover_j x \ White_i x \ Post_t + \mathbf{\Gamma} \boldsymbol{X_i} + \epsilon_{ijt}. \end{aligned} \tag{2}$$

$Callback_{ijt}$ indicates whether the applicant received a callback, and $Post_t$ indicates whether the application was sent post-BTB. $Box\ Remover_j$ is an indicator (coded at the store level) for whether the store had the box before BTB and removed it after BTB. It is coded as 0 if the store never had the box, and also in the rarer case of stores that had the box and failed to remove it after BTB. $\boldsymbol{X_i}$ is a vector of control variables including GED, employment gap, and (in the full-sample analysis only) geographic center fixed effects. The main effect of interest is the triple-differences coefficient, β_7, which tells us how the employer callback gap for whites versus blacks changes differentially after BTB for box-remover versus other stores. In column (3), to account for imbalances across chains in the pre- and postperiods, we substitute chain fixed effects, interacted with $White_i$ and $Post_t$, in place of the $Box Remover_j$, $White_i \ x \ Box Remover_j$, and $Post_t \ x \ Box Remover_j$ terms in the equation above.[35]

The triple-differences estimate ($Box Remover_j \ x \ White_i \ x \ Post_t$) in every column is large: 3.9 percentage points in the balanced sample, and 4.0 or 3.5 percentage points in the full sample depending on whether the interacted chain fixed effects

35. This produces the following estimating equation: $Callback_{ijkt} = \alpha + \beta_1 White_i + \beta_2 Post_t + \sum_{i=1}^{N} \beta_{3i} Chain_k + \beta_4 White_i \ x \ Post_t + White_i \ x \ \sum_{i=1}^{N} \beta_{5i} Chain_k + Post_t \ x \ \sum_{i=1}^{N} \beta_{6i} Chain_k + \beta_7 Box Remover_j \ x \ White_i \ x \ Post_t + \epsilon_{ijkt}$ where now k indexes chains to which store j belongs, and $Chain_k$ are dummy variables for each of the chains in our sample. Note that this analysis does not provide meaningful estimates for the effects of *White* or *Post* because they are diffused across the interacted fixed effects. However, *BoxRemover x White x Post* retains the same interpretation as in equation (2) above. The smallest chains (fewer than three locations or 12 total observations) are combined into industry-category groups; these chains represent about 9% of the sample. Use of the original coding does not affect the coefficient. In addition, in the unusual cases where "box remover" status varies between stores within a chain (usually between New Jersey and New York City), we assign separate *Chain* fixed effects to box-remover and non–box-remover subsets of such chains. The result is that the *Chain* fixed effects perfectly parallel the *Box Remover* variable.

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

are included. This implies that in the balanced sample the white callback-rate advantage over identical black applicants grew by 3.9 percentage points after BTB. These samples and specifications parallel the three analyses presented in the temporal difference-in-differences results above, but with a third difference (box remover versus other employers) added. If one compares the triple-differences estimates in Table V to the corresponding Box_j *x* $White_i$ estimates in Table IV (columns (2)–(4), respectively), in each case, the triple-differences effect estimate is slightly larger in magnitude (with signs reversed, because we are now evaluating the effect of removing the box). This pattern suggests that unrelated temporal trends are not what drove the large post-BTB expansion in racial discrimination at companies that removed the box. As we saw in Table IV, column (5), no such changes were observed at companies whose job applications were unaffected by BTB—indeed, if anything, the underlying trends cut slightly in the opposite direction.

Instead, these analyses provide evidence that BTB increases racial discrimination in employer callbacks. Prior to the adoption of BTB, racial disparities are somewhat larger among the stores that do not have the box. After BTB, that difference flips. The growth in the "white" effect after BTB is quite dramatic. In column (1) (the balanced-sample analysis, which we consider our main triple-differences specification), the estimated race gap at box-remover stores goes from 0.7 percentage points before BTB to 4.7 points after. Comparing this to the baseline callback rate, whites receive 6.7% more callbacks than similar black candidates when employers are able to observe criminal records, but 45.2% more callbacks than similar black candidates when employers cannot observe records. In other words, the race gap grows by a factor of almost 7.

Despite its advantages in terms of causal identification, the triple-differences approach comes at a cost in statistical power. Three-way interactions demand much larger samples than analyses of main effects or two-way interactions do to provide equivalent power to estimate effects of a given size; hence, the corresponding Table IV estimates are more precise even though they are estimated in smaller subsamples. Even so, the triple-differences estimates in Table V are significant at the .05 threshold or very close to it, with *p*-values ranging from .029 to .052. Moreover, they confirm the patterns observed with more

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

precision in the double-differences analyses—and provide additional confidence in interpreting these estimates as causal.

Interpretation of the triple-differences estimates as causal does rely on the assumption that, absent BTB, trends in employer callback differences by race would have been the same for employers that had the box in the preperiod and those that did not. Although our data are not long enough to compare preperiod trends, we believe the assumption is very plausible. For a vast majority of employers in our sample (even those that are franchised), the job applications are standardized nationally at the chain level, with built-in variations accommodating local differences in BTB laws.[36] Thus, the decision to include or not include the box on the application is made at the chain level, whereas callback decisions are made at the individual store level by store managers, or in some chains by local managers who supervise a small subset of locations. In that sense, whether a store has the box should be exogenous to the decision makers we are studying. Thus, we consider it appropriate to consider "box remover" status as a "treatment" variable; from the perspective of the local decision makers, it is something that is imposed on them, rather than a choice.

Moreover, there is no qualitative reason to believe that box-remover chains differ from other chains in any way that would affect hiring trends in a racially disparate way (see Online Appendix A3 for characteristics of box-remover and other employers). To pose a threat to identification in the triple-differences analysis, hiring differences across those two groups of employers would have to be racially disparate in a way that also differs over the short time between our pre- and postperiod applications (about four months on average). Note that not having the box does not generally reflect lack of interest in criminal records; chains with and without the box routinely do back-end background checks, and their applications usually warn applicants of this fact.

IV.D. Alternative Specifications and Samples

Table VI, Panels A and B show robustness checks for the balanced-sample estimates of BTB's effects on racial discrimination. Panel A shows alternative estimates for the $Box_j \times White_i$ coefficient from the temporal difference-in-differences analysis; the corresponding main specification result

36. For such chains, applications that normally have the box will usually either omit it if the store being applied to is in a BTB jurisdiction, or instruct the applicant not to answer the question if applying in certain jurisdictions.

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

TABLE VI

EFFECTS OF BAN THE BOX ON RACIAL DISCRIMINATION: ROBUSTNESS CHECKS, BALANCED SAMPLE

	Main (1)	Interview (2)	Add rev. compl. (3)	Drop RA errors (4)	NJ only (5)	NYC only (6)	Intent-to-treat (7)	Retail int. (8)
Panel A: Temporal difference-in-differences								
Box × white	−0.036**	−0.041**	−0.034**	−0.035**	−0.050*	−0.029	−0.027**	
(*Pre* × *white, column* (7))	(0.014)	(0.016)	(0.013)	(0.013)	(0.028)	(0.019)	(0.013)	
N	3,712	3,712	3,848	3,686	1,400	2,312	4,112	
Panel B: Triple differences								
Box Remover × post × white	0.039*	0.039**	0.034*	0.038*	0.048	0.034	0.026	0.041*
(*Preperiod box* × *post* × *white, column* (7))	(0.020)	(0.019)	(0.018)	(0.020)	(0.040)	(0.021)	(0.020)	(0.023)
N	11,188	11,188	11,324	11,160	4,376	6,812	11,188	11,188

Notes. Standard errors clustered on the chain in parentheses. Panel A shows robustness checks corresponding to the temporal difference-in-differences specification from Table IV, column (2); Panel B shows robustness checks corresponding to the triple-differences specification from Table V, column (1). All analyses shown are carried out in the balanced sample, consisting of stores to which we sent two applications each in the pre-BTB and post-BTB periods. Except as described below, the Panel A regressions are further confined to "box removers" (employers that had the box before BTB and removed it afterward), whereas the Panel B regressions include non–box-removers, which provide the third difference. In each panel, column (1) shows the coefficient of interest (*Box* × *white* or *BoxRemover* × *Post* × *White*, respectively) from the respective main specification. The remaining columns show modifications of the sample or specification. Column (2) uses the interview as the dependent variable (all others use "callback"). Column (3) adds observations ("reverse compliers") that had the box in the post-BTB period but not in the preperiod; in the triple-differences analysis, these are coded −1 on the *Box Remover* variable. Column (4) drops instances where RAs erred and answered a box question they were not required to answer or did not answer one they should have. Columns (5) and (6) are restricted to the New Jersey and New York City subsets of the sample, respectively. Column (7) shows an "intent to treat"–style analysis that groups companies that had the box but failed to omit it after BTB (noncompliers) with the box removers rather than with non–box-removers. In both panels, this requires changing the variable of interest and its interactions: in Panel A *Pre* (a pre-BTB period indicator) is substituted for *Box* and the coefficient shown is *Pre* × *White*; and in Panel B *PrePeriodBox* (an indicator for whether the store had the box before BTB) is substituted for *Box Remover*, and the coefficient shown is *Pre period box* × *Post* × *White*. In column (8) (Panel B only), a *Retail* industry indicator (interacted with *Post*, *White,* and *Post* × *White*) is added to the regression. *10%, **5%, and ***1% significance level.

From callbacks to labor markets - Agan & Starr, Ban the Box (2018)

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

comes from Table IV, column (2). Panel B shows alternative estimates for the $BoxRemover_j$ x $White_i$ x $Post_t$ coefficient from the triple-differences analysis; the corresponding main specification result comes from Table V, column (1). The main specification coefficient is shown in column (1) of each panel for comparison purposes. The two sets of results shown complement one another, one more precise and the other addressing additional causal inference challenges. In every specification and sample, point estimates indicate an economically large (ranging from 2.6 percentage points to 5.0 percentage points) post-BTB increase in racial discrimination. The effect sizes and precision are similar to the main specifications for most variants (with p-values close to .01 for the double-differences and .05 for the triple-differences), with exceptions discussed below.

Column (2) in both panels replaces the callback outcome variable with another variable called "interview": whether an employer's message specifically mentioned an interview. In percentage-point terms, the estimates are similar, but because baseline "interview" rates were much lower (7.5% in the Panel A sample and 6.3% in the Panel B sample) than the corresponding callback rates, BTB's apparent effect on "interview" disparities was considerably larger in proportional terms. However, we believe the interview variable is not really a good measure of whether an interview is sought (whether the employer happened to say that word generally seemed arbitrary), and that "callback" is thus the better measure.

Columns (3) and (4) in both panels alter choices that we made about how to deal with small groups of "problem" observations. Column (3) adds a group we had excluded: "reverse complier" stores that had no box before BTB but (apparently due to administrative mistakes) added it after BTB. *Box Remover* cannot be coded as 0 or 1 for these observations, but in Panel B, column (3), we code it as −1, reflecting the reversal of the usual treatment direction.[37] The effect sizes in both panels are smaller, but only slightly. In column (4), we exclude a small number of observations (about 0.3% of the full balanced sample and 0.7% of the box-remover sample) in which an RA mistakenly answered a

37. The relationship between treatment and the passage of time is inverted for these observations, making these specifications diverge from a standard difference-in-differences analysis. This is the primary reason we excluded them from both sets of the main analyses.

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

"box" question that she was not required to answer, or vice versa.[38] Excluding them leaves both estimates virtually unchanged.

Columns (5) and (6) divide the sample between New Jersey and New York City, respectively. The large reduction in sample size renders these analyses underpowered for the purpose of estimating triple differences (or even two-way interactions confined to the box-remover subset of each sample), and thus these estimates are quite imprecise and should not be given much interpretive weight. In any event, although the New Jersey point estimates are larger in percentage-point terms, they are very similar in relative terms, once one accounts for New Jersey's substantially higher callback rate (14.7% versus 9.0% in New York City, in the balanced samples).

In our main analyses, we treated the companies that retained the box after BTB (noncompliers) as part of the non–box-remover control group. We consider this to be the most appropriate categorization, because their applications did not change and hiring managers continued to be informed of criminal records, so one should expect no BTB-driven changes in racial discrimination. Moreover, the failure to comply clearly appears to be the result of choices (or, almost certainly, administrative mistakes) made at chains' national headquarters. Because this is effectively exogenous from the perspective of the local managers whose decisions are being studied, we do not particularly worry about noncompliance introducing treatment selection bias. If one's interest is in BTB's net market effects on box employers, noncompliance might reasonably be viewed as an offsetting component of those effects—albeit in most cases a temporary component, lasting until employers discover the mistake.[39] But we are more interested conceptually in understanding how access to criminal record information affects hiring managers' use of race-based assumptions about records and their resulting willingness to racially discriminate. This access did not change for noncompliant chains, so we consider them untreated observations.

38. The main sample kept RA-error cases if the same error was made consistently within the store; we coded *Box Remover* according to how the RA interpreted the application, since that tracked the information about criminal records that the RA provided to the employer.

39. We have been able to recheck most noncomplier applications, confirming that as of February 2017, most have complied with BTB.

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

Nonetheless, for readers who are more interested in BTB's effect on all preperiod box employers in the aggregate, in column (7), we take an alternate approach, treating the noncompliers as equivalent to box removers—effectively, an intent-to-treat analysis. Thus, in Panel A, we add them to the sample and replace the *Box* and *Box* × *White* variables with *Pre* and *Pre* × *White* (evaluating the effect of the passage of time on businesses that initially had the box, regardless of whether they actually removed the box). In Panel B, we replace the "box-remover" treatment variable with a "preperiod box" variable, and change the interaction terms accordingly. These changes reduce the magnitude of the point estimates (to 2.7 percentage points in Panel A and 2.6 percentage points in Panel B); they remain economically quite large, and statistically significant in Panel A but lose statistical significance in the triple differences specification in Panel B. This change is not surprising: recoding 400 noncomplier observations as though they were box removers naturally attenuates the estimates of the effects of box removal toward zero.

In column (8) of Panel B, we address a concern implicating only the triple-differences analysis. While the "box-remover" employers were qualitatively similar across numerous dimensions to the control-group employers (Online Appendix A3), they did include a larger share of retail and lower share of restaurant employers. One might worry that these industries experienced different time trends (affecting black and white applicants differently), and that this might explain our results. Although we know of no specific such trend to worry about, to test the general theory, we added an indicator for *Retail*, interacted with *White*, *Post*, and *Post* × *White*. Column (8) shows that the *Box Remover* × *Post* × *White* coefficient is if anything slightly larger when these terms are added, albeit less precise. Similar analyses (not shown here) demonstrate that the other employer characteristics described in Online Appendix A3 do not explain the triple-differences effect either.[40]

40. We also tested alternative clusterings of standard errors. The clustering shown in all tables is on the chain, because whole chains are likely susceptible to serially correlated shocks. The chain also encompasses the smaller units according to which the applications we sent were grouped (the store, or sometimes a small group of stores). If one clusters on the geographic center instead, the *p*-values for our main specifications are easily below .05 for both analyses, .019 for the temporal difference-in-differences analysis, and .027 for the triple-differences analysis. If one clusters on the individual store, they are .033 and .056, respectively.

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

TABLE VII

EFFECTS OF BTB ON CALLBACKS OF GED HOLDERS VERSUS HS GRADS: TRIPLE DIFFERENCES

	(1)	(2)	(3)
Box Remover × post × GED	−0.012	−0.010	−0.002
	(0.025)	(0.019)	(0.019)
Post × GED	0.021*	0.009	−0.000
	(0.013)	(0.010)	(0.009)
Box Remover × post	0.006	0.014	
	(0.030)	(0.024)	
Box Remover × GED	0.021	0.021	
	(0.022)	(0.015)	
Box Remover	−0.003	−0.012	
	(0.029)	(0.026)	
GED	−0.019*	−0.012	0.401***
	(0.011)	(0.008)	(0.129)
Post	0.004	0.005	0.473***
	(0.016)	(0.015)	(0.174)
N	11,188	14,637	14,637
Controls	Yes	Yes	Yes
Center FE	No	Yes	Yes
Chain FE	No	No	Yes
Post × chain FE	No	No	Yes
White × chain FE	No	No	Yes
Sample	Balanced	Full	Full

Notes. This table recreates Table V, substituting GED for White. Standard errors in parentheses clustered on chain. Dependent variable is whether the application received a callback. The balanced sample consists of stores to which we sent exactly four applications, one white/black pair in each period. Fixed effects can include chain, post × chain, white × chain, or center, and are included as indicated. *10%, **5%, and ***1% significance level.

We also varied whether our applicants had GEDs or high school diplomas and whether they had a one-year employment gap or not. These characteristics are also correlated with criminal records in the real world, so one might expect the weight that employers place on them to also increase after BTB. In fact, as we show later, having a GED instead of a regular diploma is a much stronger predictor of whether one has a criminal record than race is (and moreover, in contrast to race, GED status is a lawful basis for employer decision making). Thus, one might expect that after BTB, rational employers who seek to avoid interviewing people with records would rely heavily on diploma type as a proxy. Accordingly, in Table VII we recreate the analysis in Table V substituting in the GED variable for *White*. The point estimates are negative, but small and insignificant; they are

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

virtually zero in the full-sample analyses (Table VII, column (3)). So, we cannot characterize this as even suggestive evidence that employers are using GED as a proxy for criminal records after BTB. In Online Appendix A7 we perform a similar analysis with the employment gap variable. Though the point estimates are nontrivial (around 2.6 percentage points; Table A7.2) and cut in the expected direction, these estimates are statistically insignificant.

V. Discussion

V.A. Who Gains and Loses from BTB?

Our results produce mixed implications. On the one hand, they confirm BTB's basic premise: having a record poses an obstacle to employment. When employers had the box, applicants without records received 62% more callbacks than identical applicants with records did, even though those records entailed just one conviction for a minor, nonviolent felony, more than two years prior. This finding is consistent with prior auditing studies (Pager 2003; Pager, Bonikowski, and Western 2009), but it is useful to confirm it in a newer, much larger sample and a setting (online job applications) that is central to the modern job market. The practical effect of the criminal record penalty could be mitigated by the fact that most employers had no box even before BTB. But absent BTB, employers may ask about records at interviews and check records at any time, so employers' disfavoring of applicants with records may matter even without the box itself.

However, our findings also show that BTB increases racial discrimination. At "box remover" companies, the black-white gap increased sixfold: white applicants received 7% more callbacks than similar black applicants before BTB, and 43% more after BTB. Differencing out trends among nonbox employers only strengthens this conclusion, increasing the estimated growth in the gap slightly. Black applicants saw their callback rates fall by two percentage points after BTB, while white applicants saw theirs rise by two percentage points.

More specifically, as one would expect, BTB's main negative impact appears to fall on black applicants without records, while it is mainly white applicants with records whose callback rates go up. Figure II shows that after BTB, at box-remover employers in the balanced sample, callback rates for black applicants with records increased from 8.0% to 10.3%; for white applicants

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

with records they increased from 8.7% to 14.8%. Callback rates for black applicants without records decreased from 13.4% to 10.3%; for white applicants without records they actually increased from 14.1% to 14.8%.[41] However, overall callback rates increased in our whole sample over this time period, making this picture look overly rosy across the board. If we subtract out the 1.6 percentage point increase that occurred at companies whose applications were unaffected by BTB, white applicants without records see a small decrease from BTB (−0.6 percentage point) and blacks with records see a small gain (+0.7 percentage point), but blacks without records see very large losses (−5.1 percentage points) and whites with records see very large gains (+4.5 percentage points).[42]

V.B. Mechanisms: Statistical Discrimination versus Stereotyping

Our results imply that after BTB, employers use race to proxy for convictions, increasing racial discrimination. Are these employer assumptions empirically accurate, or are they relying on stereotypes about black criminality? Even accurate statistical discrimination is illegal, and conflicts with the policy objective of reducing employment disparities—but policy makers and scholars might still be interested in disentangling these mechanisms of discrimination. To make progress on this task requires outside data and some assumptions about the employer decision process.

It would not be surprising if employers made assumptions about black applicants' likely criminal records, even if those assumptions are not well founded. Lab experiments have consistently found that most Americans subconsciously associate blackness and criminality (see, e.g., Eberhardt et al. 2004;

41. For the post-BTB period, these calculations apply the same averages to those with and without records because, in our experiment, these applicants became indistinguishable once employers no longer asked about records.

42. Because the overall rise in callbacks appears unrelated to BTB and should be filtered out, it is misleading to suggest, as Emsellem and Avery (2016) do (relying on our full-sample results), that our study shows that BTB increased overall black callback rates. Note also that the Figure II numbers relied on here slightly understate the increase in the race gap that we found in the triple-differences regression analysis. In addition, because the primary negative effect of BTB is on black applicants without convictions, and in the real world most black men do not have felony convictions, the real-world effects on black applicants may well be worse than in our study, where half of our black applicants stood to benefit. See Section V.B for a back-of-the-envelope calculation.

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

Nosek et al. 2007). Bordalo et al. (2016) offer a general theoretical model for how generalizations based on a "kernel of truth" (such as somewhat higher black conviction rates) may become greatly exaggerated in the eyes of decision makers. They posit that decision makers save cognitive resources by relying on heuristics—representative types—that they then use to predict characteristics of interest. In our context, their theory implies that when hiring managers formulate expectations about black candidates, they overweight traits disproportionately represented in black populations, that is, felony convictions—even if the actual difference is small.

In assessing the accuracy of employer priors, a threshold challenge is to determine the real-world distribution of felony convictions in our relevant population. There are no comprehensive national data on the number of people with felony convictions overall or by subgroups (Bucknor and Barber 2016), but none of the data sources that are available vindicate the employer behavior we observe. For example, the National Longitudinal Survey of Youth 1997 (NLSY97) contains self-reported information about convictions at particular ages, and about other characteristics like race, education, and work history. Based on our calculations, these data show that by our applicants' age (22), 15.7% of non-Hispanic white males and 18.5% of black males have received adult criminal convictions. This gap, quite small to begin with (and much smaller than analogous race gaps in arrest and incarceration rates), disappears entirely once differences in educational attainment are accounted for. Because education is observable to employers and randomized in our study, if the NLSY97 data are representative, well-informed employers should not have assumed that our black applicants had higher conviction rates than our white applicants at all.

However, some researchers have critiqued the accuracy of self-report studies like the NLYS97, including as a tool for estimating race gaps (see Piquero and Brame 2008, for a review of criticisms). Moreover, the NLYS97 does not distinguish felonies from misdemeanors. Other data sources have drawn on correctional data to estimate felony conviction rates by race. Shannon et al. (2011) estimate that among adults of all ages and genders, 6% of non–African-Americans and 25% of African Americans have felony convictions. Bucknor and Barber (2016) estimate that among men of all ages, approximately 10% of white men and 42% of black men have felony convictions. Although these gaps

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

are much larger than those observed in the NLSY97, they are not broken down by age or other characteristics, and it is likely they would not be nearly as stark if limited to the subset paralleling our applicants: young men with only high-school-level education.[43]

Keeping in mind this wide range of estimates, we turn to the question of what our data can tell us about employer priors about black and white criminality. To do so, we employ a very simple model of employer decision making, outlined in detail in Online Appendix A8, in which the probability of an interview is linear in the perceived probability that an applicant has a criminal record:

$$\begin{aligned} \mathrm{P}(call_{ij}|black_i,\ no\ box_j) &= g_b\ \mathrm{P}(call_{ij}|black_i,\ crime_i) \\ &\quad + (1-g_b)\,\mathrm{P}(call_{ij}|black_i,\ no\ crime_i). \end{aligned} \tag{3}$$

That is, when there is no box, the probability that a black applicant i gets a callback from store j is a weighted average of the probability that the store would call back a black applicant with a criminal record and a black applicant without one (if that information were known). g_b is then the object of interest–the employer's estimate of the probability that a black applicant has a criminal record. An analogous equation could be written for white applicants to estimate g_w.

We use the callback rates reported in Figure II for the pre- and post-BTB periods among box-remover employers to back out g_b and g_w. Because the post-BTB period had a higher callback rate across the board, we first subtract out the 1.6 percentage points secular increase that we observed at employers whose

43. Because convictions can be accrued throughout life, the estimates for both whites and blacks (and thus, presumably, the percentage point gap between the two) would be much lower if the comparison focused on 22-year-olds. Reaves (2013) finds that the average age of a felony defendant is 32. Also, neither study estimates race breakdowns conditional on educational status or other socioeconomic characteristics, but Bucknor and Barber (2016) confirm that education and felony conviction rates are very strongly correlated: they estimate that about 64% of men without any high-school-level diploma have felony convictions (nearly 4 times the rate of male high-school graduates and 17 times the rate of men with any college). Because black men are highly overrepresented among dropouts and attend college at lower rates (see, e.g., Schott Foundation 2015), presumably the race gap in felony convictions must be substantially smaller if assessed within educational categories.

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

applications were unaffected by BTB. After this adjustment, inserting the observed callback rates into equation (3) implies that (on average) employers' priors for the probability that black and white applicants had felony convictions are 87% and 16%, respectively (a 71 percentage point difference). These priors appear sharply exaggerated, even relative to the largest estimates in the empirical literature. Even the black-white gap among males of all ages estimated by Bucknor and Barber (2016) is only about 32 percentage points. At the other extreme, the NLSY97 data suggest that rational employers should assume a 3 percentage point gap even if they ignore socioeconomic observables, or no gap if they are informed of the rates within educational categories. In Online Appendix A8, we explain what assumptions would justify equation (3). These are necessarily strong and simplified, although we also show that alternate sets of assumptions can lead us to quite similar estimated employer priors.

Moreover, even without relying on the estimates from this modeling exercise, more basic reasons suggest employers are getting it wrong. First, educational status (specifically, GED versus high school diploma) is a much stronger predictor of criminal records than race is. For example, among white and black men in the NLSY97, the chance of a conviction by age 22 was 35.8% for GED-holders and 12.1% for HSD-holders (and there is no race gap conditional on education). However, Table VII shows that employers do not change their reaction to a GED after BTB goes into effect—that is, they do not seem to use the GED as a proxy for records. No plausible model of well-informed employer decision making could explain why BTB greatly increases the effect of race and not that of a GED, suggesting that employers are instead relying on stereotypes or implicit racial bias. Second, consider the mere fact that (after adjusting for the secular rise in callbacks) the post-BTB black callback rate substantially declined compared to the rates when half of our black applicants had observable felony convictions. This suggests that post-BTB employers are assuming that considerably more than half of our black applicants have felony convictions—but by any plausible real-world measure, the appropriate assumption would be less than half, probably far less for young men with diplomas and several years of work experience.[44]

44. It is also possible that employers are assuming that black applicants have particularly serious criminal records qualitatively. While the conviction rate

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

In short, the pattern observed here is most consistent with a stereotyping model (such as that in Bordalo et al. 2016), in which small real-world differences are greatly exaggerated. Our data provide no means of testing how employers come to their seemingly incorrect assumptions, so we cannot offer a direct empirical test of Bordalo et al.'s (2016) theory of how stereotypes are formed. Alternate theories are possible, and multiple mechanisms could simultaneously contribute.[45] In any event, our data do support some form of stereotyping or bias explanation, rather than the interpretation that employers are engaging in empirically informed statistical discrimination.

V.C. Identification Challenges and Limitations

Our research design provides a strong basis for interpreting our estimates as causal. Because our black and white applicants to all employers in both periods have the same characteristics, and because our results hold when changes at businesses unaffected by BTB are filtered out, any remaining identification threats would have to come from unobserved differences that (i) affect box-remover versus other businesses differently, (ii) in ways that differ by race (among otherwise identical applicants), and (iii) differ across time periods as well. Such a difference is of course possible, but there is no obvious candidate for what it might be. This is especially so because the gap between the pre- and post-BTB periods is short, because the groups of businesses are qualitatively similar, and because we see approximately the same effect in New Jersey and New York City although their pre- and postperiods were seasonally nearly opposite.

---

data does not directly get at this possibility, it seems irrational to apply this assumption to a large share of applicants given that felony convictions of any sort are relatively infrequent by age 22. Nor do employers need to err on the side of caution at the callback stage to avoid any chance of a serious criminal record; BTB does not prevent them from eventually declining to hire after background checks.

45. For example, perhaps stereotypes are grounded in long-standing cultural biases with no empirical foundation. Or perhaps employers confuse the distribution of criminal convictions (the relevant distribution for our purposes, since convictions are what job applications ask about) with larger gaps in other outcomes like police stops. Or perhaps employers ignore the fact that race gaps are likely smaller after conditioning on other observable characteristics.

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

A potential concern is that BTB might encourage real-world applicants with records to apply to box-remover companies, affecting the competition our fictional applicants face. But such a change should affect all our applicants; there is no reason to expect it to cause employers to treat black applicants more adversely than identical whites, and even if there were, that would merely provide another mechanism by which BTB increases racial discrimination. Moreover, we think BTB probably did not substantially affect applicant pools, especially within the short period covered by our study. Many applicants likely do not know which employers have the box before they actually see it (usually on one of the final screens of the application): we ourselves could find no resources listing employers with and without the box. Applicants would also have to know about BTB and its effective date and be so discouraged by the box that they avoid applying, yet not discouraged by the fact that even post-BTB, employers conduct background checks.[46]

Our study has important limitations. Our applicants were only black and white men; dividing the sample into additional groups would have created serious statistical power concerns. We also mainly focused on chain employers in the retail and restaurant industries. These are important sectors for employment of people with records, but whether our results apply to other sectors or to smaller employers remains an open question for future research.

Perhaps the most significant limitation is that we were unable to study effects of BTB on ultimate hiring patterns, only callbacks, so we do not know whether firms avoid hiring applicants with records even after they "get their foot in the door."

46. A variant of this concern is that BTB might discourage people with records from applying to employers that never had the box. But this theory is even less likely to explain our results, which are driven almost entirely by changes among box-remover employers, not other employers (see Tables V and VI). Changes to nonbox employers' applicant pools would likely be even more subtle than changes to box employers' pools, as their applications do not change, and for most applicants there is likely no trade-off between applying to both business types. And given that these employers lack the box, many would likely not notice subtle changes in the percentage of their applicants with records. Another concern is that BTB could encourage even employers that never had the box to racially discriminate in callbacks because they know that (per BTB) they won't be able to screen out candidates with records at the interview. We do not observe such a change, however—and if anything, this would downward bias our triple-differences estimate.

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

Still, BTB is meant precisely to impact the initial stage of the hiring process (the stage at which most job applicants are filtered out), and our study speaks to those impacts. Moreover, a potential worst-of-both-worlds scenario is that BTB could have the negative consequence of excluding black applicants from callbacks, even if it does *not* have its intended positive consequence of increasing hiring of people with records. Note that, in line with our results, Doleac and Hansen (2016), using the Current Population Survey, find reduced employment of young black men in jurisdictions that adopt BTB, and (for private employers) an increase in employment of young white men.

VI. Conclusion

BTB policies are well intentioned, and they do appear to help some applicants with records to gain access to job interviews. However, our field experiment shows that this gain comes at the expense of another group that faces serious employment challenges: black men. Our findings support the hypothesis that employers substantially increase discrimination on the basis of race after BTB goes into effect.

One caveat is that our estimates (like those of auditing studies generally; see Heckman and Siegelman 1993; Heckman 1998) do not directly speak to changes in actual markets. Real-world applicants are not divided 50/50 between identical black male and white male candidates (and no other groups). And if BTB helps black men with records while hurting black men without records, the net effect on black male employment would depend on the real-world sizes of these groups.

That said, back-of-the-envelope calculations point to an enlargement of the black-white employment gap. To render these calculations conservative, we apply the conviction-rate figures from Bucknor and Barber (2016), with the largest race gap we have found in the literature. Suppose all black and white men were subject to changes paralleling the pattern in Figure II, adjusting for the 1.6 percentage point secular rise in callbacks observed at control-group companies, as described in Section V.A above. Applying these changes to the real-world distribution of records from Bucknor and Barber (2016) implies that black callback rates would fall by 2.7 points, while white callback rates would rise by 0.1 points—a net rise of 2.8 percentage points in the black-white gap. To put this in perspective, this

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

is one-quarter of our overall callback rate, and would be enough to quintuple the underlying pre-BTB black-white gap observed in our sample.[47]

Policy makers might also consider whether other interventions could offset BTB's adverse effects on black candidates. If laws against racial discrimination in hiring were effectively enforced, BTB would not have this unintended consequence.[48] This, to be sure, is easier said than done, as hiring discrimination laws are notoriously difficult to enforce. The intuition behind BTB perhaps suggests one potential innovation: employers could blind themselves to names (and other potentially racially identifying information unrelated to job qualifications).

Alternatively, policy makers could seek alternate strategies to increase hiring of people with records. One possibility is expanding successful job training and placement programs in correctional or prisoner reentry settings, though these programs have had mixed results (Visher, Winterfield, and Coggeshall 2005; Seiter and Kadela 2009; Jacobs 2012).[49] In addition, tax credits can incentivize hiring of disadvantaged groups (Katz 1998). Although a federal Work Opportunity Tax Credit (WOTC) already extends to persons hired within one year of conviction or release from prison, take-up has been limited, perhaps implying that the credit should be increased, better publicized, and/or made administratively less burdensome (Holzer, Raphael, and Stoll 2007). Other approaches

47. A similar exercise with the NLSY97 estimates leads to an estimated increase in the race gap of 4.2 percentage points. One complicating factor is that only about half of real-world applicants have racially distinctive names, perhaps reducing the relative impact of the racial-discrimination effect. However, this point may be offset by the fact that real-world applicants (unlike our fictional ones) often have other racial signals on their job applications, such as their neighborhood or high school. Moreover, even if we cut the expected losses to black and white applicants without records in half, the exercise above using Bucknor and Barber's numbers would still lead to a growth in the black-white gap of 1.4 percentage points. Of course, a full analysis would also have to consider the fact that white and black men have different distributions of other characteristics as well, and that they are not the only two groups competing for jobs.

48. Thus, we do not disagree with Emsellem and Avery (2016) that the "root of the problem" is employers' reliance on race-based assumptions about criminality; however, unless some other strategy for changing that employer behavior can be found, BTB is likely to have the unintended consequences we identify.

49. Some job training programs for disadvantaged persons more generally have been more successful (see, e.g., Roder and Elliot 2014, discussing the Year Up program) and perhaps could be adapted to target people with records.

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

could seek to mitigate employers' fear of negligent-hiring liability, which surveys show is a common reason that employers investigate applicants' records (Society for Human Resource Management 2012).[50] For example, some jurisdictions offer certificates of "employability" (or rehabilitation); these may increase employer confidence that applicants are not a substantial risk, and if litigation does arise, employers can point to the certificates to show that they took adequate care in hiring. A recent audit study found that applicants with convictions and employability certificates were called back at nearly the same rate as applicants with clean records (Leasure and Anderson 2016).

Racial disparity is not the only policy consideration surrounding BTB, and policy makers could seek to prioritize opportunities for people with records in spite of BTB's unintended racial consequences or to mitigate those consequences in other ways. But to the extent that advocates hope that BTB itself will reduce racial disparity in employment, that hope appears misguided.

RUTGERS UNIVERSITY
UNIVERSITY OF MICHIGAN

SUPPLEMENTARY MATERIAL

An Online Appendix for this article can be found at *The Quarterly Journal of Economics* online. Data and code replicating the tables and figures in this article can be found in Agan and Starr (2017b), in the Harvard Dataverse, doi:10.7910/DVN/VPHMNT.

REFERENCES

Agan, Amanda Y., and Sonja B. Starr, "Ban the Box, Criminal Records, and Statistical Discrimination: A Field Experiment," University of Michigan Law and Econ Research Paper No. 16-012 (2016), https://ssrn.com/abstract=2795795.
——, "The Effect of Criminal Records on Access to Employment," *American Economic Review*, 107 (2017a), 560–64.

50. Negligent hiring (i.e., hiring an unfit employee) is a legal theory by which employers can be held responsible for torts committed by their employees, and such suits can be costly, with settlements averaging $1.6 million (Connerly, Avery, and Bernardy 2001). Human resources guides consistently advise employers that liability is more likely if they fail to check applicants' backgrounds and to reject applicants with criminal records that suggest a risk (Connerly, Avery, and Bernardy 2001; Levashina and Campion 2009).

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

——, "Replication Data for: 'Ban the Box, Criminal Records and Statistical Discrimination: A Field Experiment,"' Harvard Dataverse (2017b), doi:10.7910/DVN/VPHMNT.
Aigner, Dennis J., and Glen G. Cain, "Statistical Theories of Discrimination in Labor Markets," *Industrial and Labor Relationship Review*, 30 (1977), 175–187.
Allen, Kathy Grannis, "NRF Survey Finds Nearly All Retailers Rely on Background Checks to Keep Consumers, Companies Safe," *National Retail Federation*, October 4, 2011, https://nrf.com/media/press-releases/nrf-survey-finds-nearly-all-retailers-rely-background-checks-keep-consumers.
Arrow, Kenneth Joseph, "The Theory of Discrimination," in *Discrimination in Labor Markets,* Orley Ashenfelter and Albert Rees, eds. (Princeton, NJ: Princeton University Press, 1973).
Astone, Nan, Michael Katz, and Juila Gelatt, "Innovations in NYC Health & Human Services Policy: Young Men's Initiative," Urban Institute, February 2014.
Baert, Stijn, and Elsy Verhofstadt, "Labour Market Discrimination Against Former Juvenile Delinquents: Evidence from a Field Experiment," *Applied Economics*, 47 (2015), 1061–1072.
Bartik, Alexander Wickman, and Scott Nelson, "Credit Reports as Resumes: The Incidence of Pre-Employment Credit Screening," MIT Department of Economics Graduate Student Research Paper 16-01 (2016).
Bertrand, Marianne, and Sendhil Mullainathan, "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," *American Economic Review*, 94 (2004), 991–1013.
Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer, "Stereotypes," *Quarterly Journal of Economics*, 131 (2016), 1753–1794.
Brame, Robert, Shawn D. Bushway, Ray Paternoster, and Michael G. Turner, "Demographic Patterns of Cumulative Arrest Prevalence by Ages 18 and 23," *Crime and Delinquency*, 60 (2014), 471–486.
Bucknor, Cherrie, and Alan Barber, "The Price We Pay: Economic Costs of Barriers to Employment for Former Prisoners and People Convicted of Felonies," CEPR Working Paper No. 2016-07 2016.
Bushway, Shawn D., "Labor Market Effects of Permitting Employer Access to Criminal History Records," *Journal of Contemporary Criminal Justice*, 20 (2004), 276–291.
Center for Health Statistics. Birth/EBC Confidential Data Files, New Jersey Department of Health, Trenton, NJ.
Clarke, H., "Protecting the Rights of Convicted Criminals: Ban the Box Act of 2012," *Washington Post*, December 20, 2012.
Clifford, Robert, and Daniel Shoag, "'No More Credit Score': Employer Credit Check Bans and Signal Substitution," Unpublished manuscript, 2016.
Color of Change, "Civil Rights Group Responds to the 'Ban the Box' Executive Order," last modified November 2, 2015, http://colorofchange.org/press/releases/2015/11/2/civil-rights-group-responds-ban-box-executive-orde/.
Community Catalyst, "Banning the Box in Minnesota—and across the United States," last modified December 2, 2013, http://www.communitycatalyst.org/blog/banning-the-box-in-minnesota-and-across-the-united-states#.UuG1_Yo46U.
Connerly, Mary L., Richard D. Avery, and Charles J. Bernardy, "Criminal Background Checks for Prospective and Current Employees: Current Practices among Municipal Agencies," *Public Personnel Management*, 30 (2001), 173–183
Decker, Scott H., Natalie Ortiz, Cassia Spohn, and Eric Hedberg, "Criminal Stigma, Race, and Ethnicity: The Consequences of Imprisonment for Employment," *Journal of Criminal Justice*, 43 (2015), 108–121.

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

Del Castillo, Christine, "Does Anyone Advertise Jobs in Newspapers Anymore," Workable (blog), May 19, 2016, https://resources.workable.com/blog/newspaper-job-ads.
Deming, David J., Noam Yuchtman, Amira Abulafi, Claudia Goldin, and Lawrence F. Katz, "The Value of Postsecondary Credentials in the Labor Market: An Experimental Study," *American Economic Review*, 106 (2016), 778–806.
Doleac, Jennifer L., and Benjamin Hansen, "Does 'Ban the Box' Help or Hurt Low-Skilled Workers? Statistical Discrimination and Employment Outcomes When Criminal Histories Are Hidden," NBER Working Paper 22469, 2016.
Eberhardt, Jennifer L., Phillip Atiba Goff, Valerie J. Purdie, and Paul G. Davies, "Seeing Black: Race, Crime, and Visual Processing," *Journal of Personality and Social Psychology*, 87 (2004), 876.
Emsellem, Maurice, and Beth Avery, "Racial Profiling in Hiring: A Critique of New 'Ban the Box' Studies," National Employment Law Project, 2016.
Fang, Hanming, and Andrea Moro, "Theories of Statistical Discrimination and Affirmative Action: A Survey," in *Handbooks of Social Economics Vol 1A*,Jess, Benhabib, Matthew O. Jackson, and Alberto Bisin, eds. (Amsterdam: North-Holland, 2011), 133–200.
Farber, Henry S., Dan Silverman, and Till Von Wachter, "Factors Determining Callbacks to Job Applications by the Unemployed: An Audit Study," NBER Working Paper 21689, 2015.
Finlay, Keith, "Stigma in the Labor Market," Unpublished manuscript, 2014.
Fryer, Roland G., and Steven D. Levitt, "The Causes and Consequences of Distinctively Black Names," *Quarterly Journal of Economics*, 119 (2004), 767–805.
Heckman, James J., "Detecting Discrimination," Journal of Economic Perspectives, 12 (1998), 101–116.
Heckman, James J., and Paul A. LaFontaine, "The American High School Graduation Rate: Trends and Levels," *Review of Economics and Statistics*, 92 (2010), 244–262.
Heckman, J., and P. Siegelman, "The Urban Institute Audit Studies: Their Methods and Findings," in *Clear and Convincing Evidence: Measurement of Discrimination in America*, Michael Fix and Raymond Struyk, eds. (Washington, DC: Urban Institute Press, 1993).
Holzer, Harry J., Steven Raphael, and Michael A. Stoll, "Perceived Criminality, Criminal Background Checks, and the Racial Hiring Practices of Employers," *Journal of Law and Economics*, 49 (2006), 451–480.
——, "The Effect of an Applicant's Criminal History on Employer Hiring Decisions and Screening Practices: Evidence from Los Angeles," in *Barriers to Reentry? The Labor Market for Released Prisons in Post-Industrial America*, Shawn Bushway, Michael Stoll, and David Weiman, eds. (New York: Russel Sage Foundation, 2007).
Jacobs, Erin, "Returning to Work after Prison—Final Results from the Transitional Jobs Reentry Demonstration," May 10, 2012, http://dx.doi.org/10.2139/ssrn.2056045.
Katz, Lawrence F. "Wage Subsidies for the Disadvantaged," in *Generating Jobs*, Richard B. Freeman and Peter Gottschalk, eds. (New York: Russel Sage Foundation, 1998).
Kroft, Kory, Fabian Lange, and Matthew J. Notowidigdo, "Duration Dependence and Labor Market Conditions: Evidence from a Field Experiment," *Quarterly Journal of Economics*, 128 (2013), 1123–1167.
Lahey, Joanna N., "Age, Women, and Hiring an Experimental Study," *Journal of Human Resources*, 43 (2008), 30–56.
Lahey, Joanna N., and Ryan A. Beasley, "Computerizing Audit Studies," *Journal of Economic Behavior and Organization*, 70 (2009), 508–514.
Leasure, Peter, and Tia Stevens Andersen, "The Effectiveness of Certificates of Relief as Collateral Consequence Relief Mechanisms: An Experimental Study," *Yale Law & Policy Review, Inter Alia*, 35 (2016).
Levashina, Julia, and Michael A. Campion, "Expected Practices in Background Checking: Review of the Human Resources Management

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

Literature," *Employee Responsibilities and Rights Journal*, 21 (2009), 231–249.
List, John A., "The Nature and Extent of Discrimination in the Marketplace: Evidence from the Field," *Quarterly Journal of Economics*, 119 (2004), 49–89.
Love, Margaret Colgate, "LAWYERING: Paying Their Debt to Society: Forgiveness, Redemption, and the Uniform Collateral Consequences of Conviction Act," *Howard Law Journal*, 54 (2011), 753–795.
Minnesota Department of Human Rights, "Ban The Box: Overview for Private Employers," accessed January 19, 2016, https://mn.gov/mdhr/employers/criminal-background/.
NAACP, "Ban the Box," accessed May 3, 2017, http://www.naacp.org/campaigns/ban-the-box/.
Neumark, David, "Detecting Discrimination in Audit and Correspondence Studies," *Journal of Human Resources*, 47 (2012), 1128–1157.
Neumark, David, Roy J. Bank, and Kyle D. Van Nort, "Sex Discrimination in Restaurant Hiring: An Audit Study," *Quarterly Journal of Economics*, 111 (1996), 915–941.
Neumark, David, Ian Burn, and Patrick Button, "Is It Harder for Older Workers to Find Jobs? New and Improved Evidence from a Field Experiment," NBER Working Paper 21669, 2015.
Nosek, Brian A., Frederick L. Smyth, Jeffrey J. Hansen, Thierry Devos, Nicole M. Lindner, Kate A. Ranganath, Colin Tucker Smith, Kristina R. Olson, Dolly Chugh, Anthongy G. Greenwald, and Mahzarin R. Banaji, "Pervasiveness and Correlates of Implicit Attitudes and Stereotypes," *European Review of Social Psychology*, 18 (2007), 36–88.
Oreopoulos, Philip, "Why Do Skilled Immigrants Struggle in the Labor Market? A Field Experiment with Thirteen Thousand Resumes," *American Economic Journal: Economic Policy*, 3 (2011), 148–171.
Pager, Devah, "The Mark of a Criminal Record," *American Journal of Sociology*, 108 (2003), 937–975.
Pager, Devah, Bart Bonikowski, and Bruce Western, "Discrimination in a Low-Wage Labor Market: A Field Experiment," *American Sociological Review*, 74 (2009), 777–799.
Phelps, Edmund S., "The Statistical Theory of Racism and Sexism," *American Economic Review*, 62 (1972), 659–661.
Pinard, M., "Ban the Box in Baltimore," *Baltimore Sun*, January 7, 2014.
Piquero, Alex R., and Robert W. Brame, "Assessing the Race–Crime and Ethnicity–Crime Relationship in a Sample of Serious Adolescent Delinquents," *Crime and Delinquency*, 54 (2008), 390–422.
Reaves, Brian A., "Felony Defendants in Large Urban Counties, 2009-Statistical Tables," US Department of Justice, 2013.
Riach, Peter A., and Judith Rich, "Field Experiments of Discrimination in the Market Place," *Economic Journal*, 112 (2002), F480–F513.
Roder, Anne, and Mark Elliott, "Sustained Gains: Year Up's Continued Impact on Young Adults' Earnings," Economic Mobility Corporation Report (May 2014), https://economicmobilitycorp.org/uploads/sustained-gains-economic-mobility-corp.pdf.
Rodriguez, Michelle N., and Beth Avery, "Ban the Box: US Cities, Counties, and States Adopt Fair Hiring Policies," National Employment Law Project (2016), accessed June 9, 2016, http://www.nelp.org/publication/ban-the-box-fair-chance-hiring-state-and-local-guide/.
Scarborough, David, "Does Job Testing Harm Minority Workers? Evidence from Retail Establishments," *Quarterly Journal of Economics*, 123 (2008), 219–277.
Schott Foundation, "Black Lives Matter: Schott 50-State Report on Public Education and Black Males," 2015, http://blackboysreport.org/national-summary/.
Seiter, Richard P., and Karen R. Kadela, "Prisoner Reentry: What Works, What Does Not, and What Is Promising," *Crime and Delinquency*, 49 (2009), 360–388.

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

Shannon, Sarah, Christopher Uggen, Melissa Thompson, Jason Schnittker, and Michael Massoglia, "Growth in the U.S. Ex-Felon and Ex-Prisoner Population, 1948–2010," Unpublished manuscript, 2011.
Shoag, Daniel, and Stan Veuger, "No Woman No Crime: Ban the Box, Employment, and Upskilling," Unpublished manuscript, 2016.
Society for Human Resource Management, "Background Checking—The Use of Criminal Background Checks in Hiring Decisions," last modified July 19, 2012, https://www.shrm.org/hr-today/trends-and-forecasting/research-and-surveys/pages/criminalbackgroundcheck.aspx.
Southern Coalition for Social Justice, "Ban the Box Community Initiative Guide," accessed May 3, 2017, http://www.southerncoalition.org/program-areas/criminal-justice/ban-the-box-community-initiative-guide/.
Stoll, Michael A., "Ex-Offenders, Criminal Background Checks, and Racial Consequences in the Labor Market," *University of Chicago Legal Forum* (2009), Article 11.
Uggen, Christopher, Mike Vuolo, Sarah Lageson, Ebony Ruhland, and Hilary K. Whitham, "The Edge of Stigma: An Experimental Audit of the Effects of Low-level Criminal Records on Employment," *Criminology*, 52 (2014), 627–654.
U.S. Bureau of Labor Statistics, "Labor Force Characteristics by Race and Ethnicity, 2015," U.S. Department of Labor, September 2016, https://www.bls.gov/opub/reports/race-and-ethnicity/2015/home.htm.
U.S. Equal Employment Opportunity Commission (EEOC), "Consideration of Arrest and Conviction Records in Employment Decisions under Title VII of the Civil Rights Act of 1964," EEOC Enforcement Guidance, April 25, 2012, https://www.eeoc.gov/laws/guidance/arrest_conviction.cfm.
Visher, Christy A., Laura Winterfield, and Mark B. Coggeshall, "Ex-Offender Employment Programs and Recidivism: A Meta-Analysis," *Journal of Experimental Criminology*, 1 (2005), 295–316.
Wozniak, Abigail, "Discrimination and the Effects of Drug Testing on Black Employment," *Review of Economics and Statistics*, 97 (2015), 548–566.

Downloaded from https://academic.oup.com/qje/article-abstract/133/1/191/4060073 by Harvard Library user on 25 October 2019

E-lit-012: Race Discrimination: An Economic Perspective

Kevin Lang and Ariella Kahn-Lang Spitzer | Journal of Economic Perspectives 34(2): 68-89 | 2020-05
https://www.aeaweb.org/articles?id=10.1257/jep.34.2.68

About this source

The clearest short statement of the sophisticated middle position, and the right closing source for what a callback gap actually licenses you to conclude. Lang and Kahn-Lang Spitzer accept that the experimental evidence of discrimination in hiring is substantial while insisting that economics still knows very little about how much of the Black-white gap in wages and employment discrimination accounts for, once pre-market skill differences are handled honestly. They supply the theoretical vocabulary the rest of the section needs - taste-based discrimination, accurate statistical discrimination, and statistical discrimination on inaccurate beliefs, which have different welfare and policy implications and are not distinguished by a callback gap - and they point out that in search and matching models a firm-level contact gap need not generate a market-level wage gap at all, since rejected applicants apply elsewhere. They also frame the economist-sociologist split over what the audit literature has established, which this briefing has to navigate rather than pick a side in. Read against the other three sources, it is the check on over-claiming: the firm-level evidence has gotten much sharper since 2020, but the aggregate accounting question they pose is still largely open. Bound in as a partial retypeset: only the published abstract, the JEL codes and the citation could be retrieved, so this paper may be cited for its stated position but no page or passage may be quoted from it. The published abstract, verbatim (the only text retrievable, via the RePEc record): "We review the empirical literature in economics on discrimination in the labor market and criminal justice system, focusing primarily on discrimination by race. We then discuss theoretical models of taste-based discrimination, particularly models of frictional labor markets and models of statistical discrimination, including recent work on invalid statistical discrimination. We explore and evaluate the evidence for and against these theories. Although there is substantial evidence of the existence of discrimination, little is known about the extent to which disparities are driven by discrimination. Finally, we argue that economists miss the important self-enforcing relationship between disparities and discrimination and the effect of disparities in one domain on discrimination in other domains."

Bound as the web text, retypeset. Retrieved 2026-09-10. sha256 df39ac38325dddac...
Bound into an Economic Briefing · by Claude, may contain mistakes

Full document begins on the next page.

S4–4: Race Discrimination: An Economic Perspective

Kevin Lang and Ariella Kahn-Lang Spitzer | Journal of Economic Perspectives 34(2): 68-89 | 2020-05

https://www.aeaweb.org/articles?id=10.1257/jep.34.2.68

Full content unavailable - Journal of Economic Perspectives full text is free to read but every AEA/Atypon PDF route returns HTTP 403 to scripted and headless-browser fetches; no repository or working-paper manuscript exists. Published abstract, JEL codes and citation captured.. Only partial content is included.

Published abstract (verbatim, Journal of Economic Perspectives 34(2), Spring 2020, pp. 68-89):

"We review the empirical literature in economics on discrimination in the labor market and criminal justice system, focusing primarily on discrimination by race. We then discuss theoretical models of taste-based discrimination, particularly models of frictional labor markets and models of statistical discrimination, including recent work on invalid statistical discrimination. We explore and evaluate the evidence for and against these theories. Although there is substantial evidence of the existence of discrimination, little is known about the extent to which disparities are driven by discrimination. Finally, we argue that economists miss the important self-enforcing relationship between disparities and discrimination and the effect of disparities in one domain on discrimination in other domains."

JEL codes: J15 (Economics of Minorities, Races, Indigenous Peoples, and Immigrants; Non-labor Discrimination); J71 (Labor Discrimination; Hiring and Firing); K42 (Illegal Behavior and the Enforcement of Law).

Citation: Kevin Lang and Ariella Kahn-Lang Spitzer, 2020. "Race Discrimination: An Economic Perspective," Journal of Economic Perspectives, American Economic Association, vol. 34(2), pages 68-89, Spring. DOI: 10.1257/jep.34.2.68.

What is and is not in this file. Only the published abstract, the JEL codes and the citation above were retrievable. The Journal of Economic Perspectives is free to read on the American Economic Association's own site, but every AEA and Atypon route to the article PDF (aeaweb.org/articles/pdf/doi/..., pubs.aeaweb.org/doi/pdf/..., pubs.aeaweb.org/doi/pdfplus/...) returned HTTP 403 to both a scripted request and a headless-browser print, and the JSTOR copy is behind a login. Unpaywall and OpenAlex both resolve the article to those same blocked AEA URLs, Semantic Scholar records it as GOLD open access but returns only the DOI, and no NBER, RePEc or institutional-repository manuscript exists: the RePEc working-paper record for this title points back to the AEA page, and neither author's personal page hosts a loose copy

(Ariella Spitzer's research page links to pubs.aeaweb.org). The full text should be quoted only from the abstract above unless a human opens the article at https://www.aeaweb.org/articles?id=10.1257/jep.34.2.68 (free, no subscription needed) and drops the PDF in.

Source: https://www.aeaweb.org/articles?id=10.1257/jep.34.2.68

E-lit-016: Investigating Prisoner Reentry: The Impact of Conviction Status on the Employment Prospects of Young Men

Devah Pager and Bruce Western | National Institute of Justice final report 228584 (the New York City in-person audit published as Pager, Western and Bonikowski, 'Discrimination in a Low-Wage Labor Market: A Field Experiment,' American Sociological Review 74(5): 777-799) | 2009

https://www.ojp.gov/pdffiles1/nij/grants/228584.pdf

About this source

The counterweight to the claim that no audit design has followed an application past the callback. Pager and Western sent matched teams of trained live testers, equivalent on every resume characteristic and differing only in race (and, in the criminal-record arm, in a randomly assigned felony conviction), through the entry-level application process at hundreds of New York City employers, and recorded job offers as well as callbacks. The in-person audit is the design Heckman and Siegelman originally attacked, and it is the only design that observes the outcome the correspondence literature cannot: whether an applicant is actually hired. Acquired as the top-up for the critic's missing-counterevidence finding, to date the callbacks-to-jobs question properly and to correct the briefing's framing that the post-callback stage is untested.

Bound as the original PDF. Retrieved 2026-09-10. sha256 69a15d9f588873d7...

Bound into an Economic Briefing · by Claude, may contain mistakes

Full document begins on the next page.

E-lit-016: https://www.ojp.gov/pdffiles1/nij/grants/228584.pdf

The author(s) shown below used Federal funds provided by the U.S. Department of Justice and prepared the following final report:

Document Title: **Investigating Prisoner Reentry: The Impact of Conviction Status on the Employment Prospects of Young Men**

Author: **Devah Pager and Bruce Western**

Document No.: **228584**

Date Received: **October 2009**

Award Number: **2005-IJ-CX-0019**

This report has not been published by the U.S. Department of Justice. To provide better customer service, NCJRS has made this Federally-funded grant final report available electronically in addition to traditional paper copies.

Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

National Institute of Justice Grant: Final Report

Title: Investigating Prisoner Reentry: The Impact of Conviction Status on the Employment Prospects of Young Men

Principle Investigators: Devah Pager and Bruce Western

Abstract:
Our research addresses the problem of prisoner reentry by focusing on the employment of ex-offenders shortly after release. We completed a three-part project to study the employment barriers facing men with criminal records. The empirical core of the project is a randomized field experiment that sent matched teams of testers to apply for hundreds of entry-level jobs in New York City. The experiment observed how employers respond to applicants who are equally qualified but vary by race, ethnicity, and criminal record. Because the research design allowed us to finely control characteristics of job seekers, and randomly assign criminal records and job openings, the experimental data yields unusually clear and convincing evidence of the impact of a criminal record. The audit study alone would provide a powerful method for studying the employment barriers to ex-offenders, but it is complemented here by a telephone survey of the same employers, and in-depth qualitative interviews with an additional subset. Combining experimental measures with interview data offers a unique opportunity to study the hiring process from both the job seeker's and the employer's point of view. Policies supporting prisoner reentry and ex-offender employment are directly informed by the rich data resulting from our multi-stage investigation.

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Title: Investigating Prisoner Reentry:
The Impact of Conviction Status on the Employment Prospects of Young Men

As record numbers of young men leave prisons, nearly 700,000 this year alone, the question of prisoner reentry has become a pressing concern. In particular, the employment outcomes of ex-offenders represents a critical issue, given the connections between stable work and desistance from crime (Uggen, 2000). Unfortunately, current statistics do not present an optimistic picture. Surveys of former inmates suggest that 60 to 75 percent remain jobless up to a year after release (Travis, 2005; Petersilia, 2003). In light of these outcomes, the fact that nearly half return to prison within three years is of little surprise (Bureau of Justice Statistics, 2002).

Adding to the general problems of prisoner reentry, the question of race presents additional complexities in the reentry process. Blacks now represent over 40 percent of the prison population (relative to just 12 percent of the general population), and are likewise significantly overrepresented among reentering inmates. In addition to higher rates of criminal records among African Americans, some evidence suggests that black ex-offenders may also pay a higher penalty for having a criminal record relative to otherwise similar whites (Pager, 2007). Given these patterns, understanding the nature of criminal and racial stigma—and the combination of the two—represents and important goal for research and policy.

In this report we present the results of a two-part study investigating the barriers of race and criminal background in access to low wage labor markets. The first stage of our research involved a large-scale field experiment in which we sent teams of trained young men to apply for entry-level jobs, varying only according to their race and criminal background. The field experiment was then followed by survey interviews and in-depth interviews with the same employers to provide more detailed information about their businesses and their concerns about hiring applicants with criminal records. This multi-method approach provides a richer perspective on the process through which employers express and enact their preferences. The

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

experimental design allows us to isolate the causal influence of a criminal record, apart from the many possible sources of spuriousness; the survey methods allow us to compare our findings with traditional measures of employer behavior; and the in-depth interviews allow us to investigate the meanings employers place on conviction status as a signal of employability. In the remainder of this report, we discuss the central findings of this research and their implications for policy on prisoner reentry and ex-offender employment.

Stage I: Hiring Experiment

The New York City Hiring Discrimination Study sent matched teams of testers to apply for hundreds of real entry-level jobs throughout New York City over nine months in 2004. The testers were well-spoken, clean-shaven young men, aged 22 to 26. Most were college-educated, between 5 feet 10 inches and 6 feet in height, recruited in and around New York City. They were matched on the basis of their verbal skills, interactional styles (level of eye-contact, demeanor, and verbosity), and physical attractiveness. Testers were assigned fictitious resumes indicating identical educational attainment, and comparable quality of high school, work experience (quantity and kind), and neighborhood of residence. Resumes were prepared in different fonts and formats and randomly varied across testers, with each resume used by testers from each race group. Testers presented themselves as high school graduates with steady work experience in entry-level jobs. Finally, the testers passed through a common training program to ensure uniform behavior in job interviews. While in the field, the testers dressed similarly and communicated with teammates by cell phone to anticipate unusual interview situations.

For the study as a whole, the ten testers were randomly assigned (and rotated) across four teams, each consisting of two or three people. In this report, we focus on the results related to the

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

effects of a criminal record on the labor market outcomes of young black and white men.[1] The teams for this part of the study each consisted of two testers of the same race; one team included two white testers, the second team included two black testers. In both teams, one tester was randomly assigned a "criminal record" for the first audit; the criminal record condition then alternated between testers for each successive employment search. Thus each tester served in the criminal record condition for an equal number of cases. By randomly allocating criminal records to testers, unmeasured differences between testers were effectively controlled. None of the testers in the study had real criminal backgrounds, but presented fictitious criminal record information to employers for the purposes of the study. The criminal record was typically disclosed in answer to the standard question on employment applications, "Have you ever been convicted of a crime? If yes, please explain." When asked, testers were instructed to reveal that they had recently been released from prison after serving 18 months for a drug felony (possession with intent to distribute, cocaine). In addition, following Pager (2003), the white tester's criminal record was additionally signaled on the resume by listing work experience at a state prison, and by listing a parole officer as a reference.[2]

Employers were sampled from job listings for entry-level positions, defined as jobs requiring no previous experience and no education greater than high school. Job titles included restaurant jobs, retail sales, warehouse work, couriers, telemarketers, customer service positions, clerical workers, stockers, movers, delivery drivers, and a wide range of other low wage positions. Job listings were randomly drawn each week from the classified sections of *The New York Times*, *The Daily News*, *The New York Post*, *The Village Voice,* and *Craigslist*. The broad

[1] A second major focus of the study was on measuring the prevalence of direct racial discrimination. Results from this part of the study appear in Pager, Western, and Bonikowski (forthcoming). For this grant, we focus primarily on outcomes relevant to the employment of ex-offenders.

[2] Results from Pager (2003) suggest that providing information about a criminal record to employers who do not request the information does little to affect hiring decisions. Those employers who request the information are those most likely to use it.

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

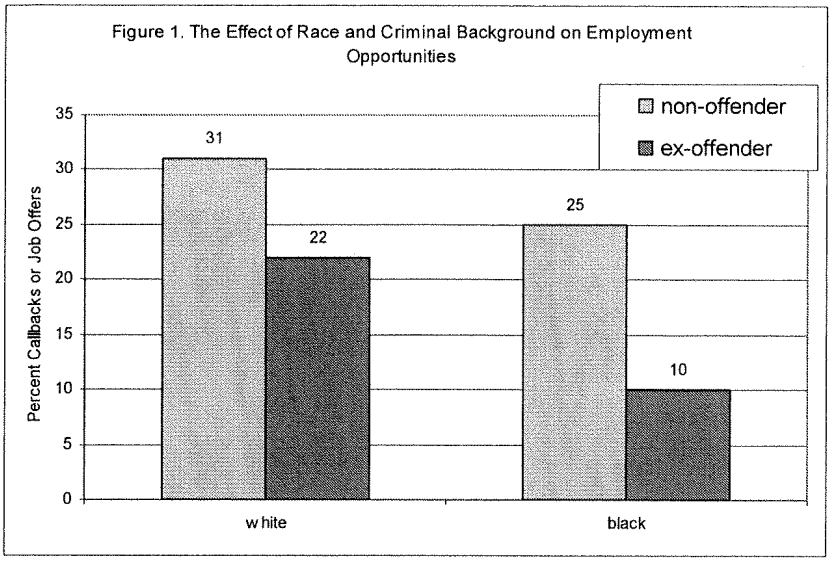
range of job listings allows for comprehensive coverage of the entry-level labor market in New York. From the available population of job listings, we took a simple random sample of advertisements each week (n = 250). Testers in each team applied to each job within a 24-hour period, randomly varying the order of the applicants.

Our dependent variable recorded any positive response in which a tester was either offered a job or called back for a second interview. Callbacks were recorded by voicemail boxes set up for each tester. For more information about the research design and methods, see Pager, Western, and Bonikowski, forthcoming.

Results

Two key findings emerge from the first set of audit results. First, not surprisingly, a criminal record has a significant negative impact on hiring outcomes, even for applicants with otherwise appealing characteristics. Across teams, a criminal record reduced the likelihood of a callback or job offer by nearly 50 percent (28% vs 15%). Second, evidence from our tester teams suggests that the negative effect of a criminal conviction is substantially larger for blacks than for whites. As shown in Figure 1, the criminal record penalty suffered by white applicants (30%) is roughly half the size of the penalty for blacks with a record (60%). The interaction between race and criminal record in these data is large and statistically significant indicating that the penalty of a criminal record is more disabling for black job seekers than whites.

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.



The intensification of the criminal record effect among blacks is consistent with earlier audit research (Pager, 2007), and points to special barriers facing blacks in the transition from prison to work. Employers, already reluctant to hire blacks, appear particularly wary of blacks with known criminal histories. In seeking to better understand the interaction between race and criminal background in shaping labor market outcomes, we looked to the sequence of interactions that lead to this ultimate pattern of results. As job applicants pass from the point of application to an interview, and from an interview to an offer, we witness some of the underlying dynamics that may shape employers' decision making and result in the systematic disadvantage of blacks with criminal records. In particular, we see ways in which opportunities for rapport-building appear to mediate the effects of race and criminal background on hiring outcomes.

The Importance of Personal Contact

One of the ways job applicants can build rapport with employers is through the interview process. Though typically brief for low wage jobs, interviews provide an opportunity for applicants to demonstrate their communication skills and their commitment to work. For employers concerned about soft skills not reflected on a resume, even a quick interaction can

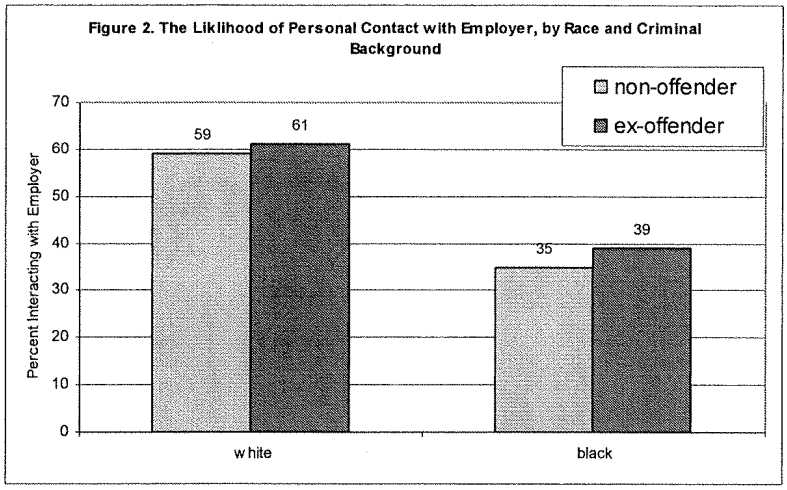
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

provide important information about the capacity of an individual applicant. Especially in the case of stigmatized applicants, personal contact may serve a particularly important purpose as a way of counteracting employers' initial stereotypes. As employers learn more about the person behind the category (e.g., ex-offender, black man), their comfort level with the applicant in question is likely to increase.

As our testers passed from the point of application to an interview, and from an interview to an offer, we witnessed some of the underlying dynamics that may shape employers' decision-making and result in the systematic disadvantage of blacks with criminal records. Indeed we find evidence that personal contact with an employer has a substantial impact on hiring outcomes. Testers who interacted with employers were nearly six times more likely to receive a callback or job offer relative to those who did not.[3] At the same time, it appears that this pathway to rapport-building may not be equally available to all applicant types. Though all testers in the study were instructed to request to speak to the person in charge of hiring and to proceed as far as they could in the interview process during their visit, some met with more success than others. In particular, race has a significant effect on the likelihood of personal contact between applicant and employer, with blacks roughly 40 percent less likely to have the opportunity to interact with employers ($p<.001$). Employers appear to be screening on the basis of race in deciding who to allow to proceed from application to interview (see Figure 2).

[3] Results from a logistic regression predicting a positive response as a function of race, criminal record, an interaction between the two, and a dummy variable for personal contact with employer. Models control for clustering on employer.

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.



By contrast, the effect of a criminal record has no discernable impact on the likelihood of interaction. Given that a criminal record is typically unobserved until an employer has spoken with the candidate and/or reviewed their materials, it is not surprising that this is not typically the basis of an employer's decision to interview.

The barriers that emerge in this very early stage of the hiring process are no doubt consequential for the ultimate disparities we observe. With fewer opportunities for face to face contact with employers, black applicants were limited in their ability to demonstrate their specific skills and attributes. Particularly in the case of black ex-offenders, for whom employers' concerns were likely to be particularly strong, limits on interaction reduced opportunities to contextualize the conviction or to demonstrate evidence of successful rehabilitation.

Quality of the Interaction

Being granted an interview, or even a conversation, with the person in charge of hiring provides an important opportunity for establishing rapport. But the nature of that rapport also hinges upon what takes place during the interaction. Particularly for applicants with criminal records, the interview provides a key opportunity to assuage employers' concerns. In this respect, our testers had mixed experiences. The testers had a set script that they were instructed to convey to

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

employers about their prior drug conviction and their commitment to rehabilitation. In some cases employers were extremely receptive to discussing these issues, while in others employers seemed uncomfortable or unwilling to broach the subject. In one interaction, for example, an employer inquired about the most recent job listed on Kevin's resumé, which was at a correctional facility. Kevin reports, "*I thought she was asking me what I did to get in there. I said, 'It was for drug possession.' She said, 'No, not that. That's none of my business.'*" The employer then quickly moved on to discuss Kevin's previous work experience. It is unclear from this interaction what lingering doubts the employer may have had about Kevin's criminal background, but Kevin does not have the chance to explain further.

In another case, Anthony, an African American tester, reports, "*As she looked over [my resumé] I could barely hear her say, 'Oh, I see.' I don't know what it was in response to, but it was pretty quick so I would guess it was my conviction... She then just looked up at me and said, 'I'll give you a call.' It seemed like she ended it a bit abruptly.*" Some employers seemed uncertain about what was legally or socially appropriate to ask about prior convictions, and others seemed simply uncomfortable with the topic or considered it outside of the realm of employment-related concerns. In these cases, it is difficult to interpret the employers' response to the criminal background, and the applicant typically has less opportunity to account for the stigmatizing record or address employers' underlying concerns.

In other cases, employers' concerns about the criminal record are more transparent. Worries about legal liability, for example, came up in this interview with Chad, an African American tester:

> *When I finished the application I was interviewed by...a large white man with a thick mustache. He shook my hand, invited me to have a seat, and began to look over my resumé. He said, "First, I need you to explain this...correctional facility and parole officer reference." I told him that I was convicted of a drug charge -- possession with intent to sell. "I can't hire you," he said. He went on to explain that a lot of things can happen and the liability is too great. He said, "Let's say you got into a fight with a guy*

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

> *and you were in the right. The police come and run your background, yes, now you're in the wrong, even though you may have been right. It wouldn't be good for you and it wouldn't be good for us. I couldn't hire you."*

Sometimes, employers' negative reactions are less explicit, but their concern over the prospect of hiring an ex-offender is clear nevertheless. For example, Kevin, a white tester, reports his experience at a specialty foods store:

> *"I noticed a sign on the door which read, 'Help wanted, part time, some experience'.... A few minutes later a man came out.... He told me that he had a great Part Time position [and] there could be some full time positions opening [in] a while. He pointed at my response for reasons for leaving on the application right next to [the correctional facility] and said, 'Why did you write parole?' I said that I was currently on parole. [He] then looked me in the eye and said, 'Did you commit a crime?' I said yes. He then looked down at the sheet and said that he really wasn't hiring right now...."*

Kevin's conviction record seemed to catch the employer by surprise. Within seconds, the many signals pointing to a job opportunity (help wanted sign, "great part time positions," etc.) disappeared, as the employer decided he was no longer hiring, or at least not hiring Kevin.

Despite these unpleasant experiences, not all employers reacted negatively upon noticing a criminal record. In fact, on a number of occasions, testers encountered extremely sympathetic employers. For example, Kevin records his experience in applying for a job at a car dealership:

> *"He saw the correctional facility [on my resumé] and said, 'We're an Equal Opportunity Employer. We don't care about this. About 75% of the people in this business have a record anyway....'"*

Kevin describes the end of the interview:

> *"He said he was going to call me. Then he said, 'I know what you are thinking. This asshole is never going to call me. I will call you. Not because you have good sales experience but just because you need a chance.'"*

This employer appears extremely sympathetic to the plight of an ex-offender looking for work. In fact, the employer seems willing to privilege the desire to give Kevin a second chance over his need for workers with relevant experience. Indeed, this employer called Kevin about the job two days later.

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

To be sure, many of the sympathetic responses received by testers in the criminal record condition simply reflect the preexisting attitudes of employers, independent of the interaction. Employers who feel sympathetic toward ex-offenders are likely to express such sympathies in conversations with ex-offender applicants. But above and beyond employers' predispositions, we observe some evidence that the interaction itself can work to clarify and shape the employers' interpretation of the criminal record. For employers who have ambivalent feelings about hiring ex-offenders, or who have anxieties about particular kinds of ex-offenders, interaction with the candidate allows the employer to interrogate these concerns directly. In one case, for example, Keith, a white tester, has an extended conversation with the manager at a furniture rental store:

> *After finishing the application I brought it back to [the employer], along with my resumé. He invited me to take off my backpack and have a seat. He began looking over the res./app. and his first question was, 'Were you selling or using?' I told him, 'Using. It was a minor thing. A stupid mistake and I'm now clean.' He was sympathetic saying, 'I gotcha. It was a question, not a criticism.' [The employer then asked him a few questions about his driver's license and driving history.] He invited me to sit with him out at the door while he smoked. When we got there he informally sat me down and lit up. He turned to me and said, 'So why should I hire you instead of one of the 47 other guys I got coming to me?' I told him, 'I'll work hard for you, bust my ass. It's a condition of my parole that I work.' He said, 'But do you want to work?' I answered, 'Yeah, I'm looking to get back into society...' He interrupted, 'You want to get your shit together.' 'Yeah,' I said. ...The conversation ended with him saying, 'My inclination is to say yes,' (regarding hiring me). He added, 'My bosses, the owners, are a little more close-minded than me.... Look, you paid, you don't have to pay for it the rest of your fucking life. People make mistakes. I'll get you my card.'*

This employer expresses some open-mindedness about Keith's criminal background from the start ("It was a question, not a criticism"), but he also wants evidence that taking a chance on Keith would be warranted amidst the large pool of other candidates. The conversation appears to provide important reassurances to the employer, who ends the interview with an encouraging note.

In another case, Anthony, an African American tester, discusses his criminal background with the manager of a healthcare company:

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

"[The employer] said, '*I'm sure people must take double-takes [when they see the correctional facility].*' I replied, '*Yeah, that does happen sometimes.*' He told me that he knows the law, says I have to provide that information to a possible employer, but not to worry because he has had other employees who have '*fucked up in the past.*' He said, *'I feel safer knowing you're telling me up front than me having to wonder if you're gonna come here and tear shit up. Let's face it, interviews are bullshit. You can't know someone from a 5-minute interview. So let's cut to the chase.'*

- '*How long were you in?*' (18 months)
- '*When'd you get out?*' (A few weeks ago)
- '*Ok, so you're fresh out and trying to get back on track?*' (Yes, I am)
- '*What'd you do?*' (I had a small amount of drugs on me)
- '*So you were guilty?*' (Yes, I was young and made some mistakes but I learned from them and am completely drug-free)
- (jokingly) '*So you weren't innocent?*' (No, it was my fault.)
- '*Don't worry, I find that those that messed up and want to set things right are better workers.*'

In this interaction the employer does acknowledge certain concerns about hiring someone with a record, but seems to respond favorably to Anthony's honesty, and after learning more of the details of Anthony's background, offers him an encouraging response.

Thus, while a criminal record does have a significant negative impact on the employment prospects of job seekers, some employers are willing to look beyond the conviction, or to downplay its significance in the context of other information they acquire during the interview. In these cases, a kind of empathy seems to develop between employer and job seeker, with goodwill often translating into a substantial improvement in employment prospects. Of course the types of individualizing information employers look to are not always in the applicants' control. Race and ethnicity, in particular, appear to affect some employers' interpretations of the seriousness of the criminal background and the depth of empathy generated by the interaction.

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

For example, Keith, a white tester, reports on his interaction with the manager of a restaurant supply company:

> *[The employer] sat me down and went over my application and resumé. He first saw [the correctional facility] and asked about my working there. I told him that I had been incarcerated. In a lower voice he said, "What did you do?" I told him about my being caught with cocaine, my time served, my current sobriety, which my parole officer could verify, the fact that it was a mistake and I had learned my lesson. At this point he said, "Zarriello...what is that, Italian?" I said, "Yes," and could immediately tell he was now on my side. He asked more questions like, "What happened exactly?" I told him I was in a car with some ex-friends that was pulled over and we were all searched and they found 6 grams in my possession. He asked, "You come from a nice Italian family? What do your parents think about this?" I told him they were disappointed but thought I had learned my lesson. He told me that "people make mistakes." He went on to say, "It would make me feel good to help a guy like you, more than just the rewards of doing my job, but good as a person if I can help someone."*

This employer's emphasis on the value of second chances and the desire to help a young ex-offender get back on his feet are similar to sympathetic reactions we heard from other employers in interactions with both black and white testers. Here, though, we see the employer explicitly invoke Keith's ethnic background as the basis for solidarity, and as a key turning point in the employer's reaction to the criminal background information, as Keith moves from being viewed as an ex-offender to someone from a "nice Italian family."

In a similar case, Kevin applied for a job with a staffing agency and was asked a number of detailed questions about his criminal history. Toward the end of the interview Kevin reports,

> *"[The employer] said, 'Do you have Irish in you?' I said, 'On my mother's side I do.' He asked what else I have. I said, 'French.' He was delighted! He said he has Irish and French in him, too. He said we could be related because we are both from FrIreland (France & Ireland). I said yes and laughed with him."* Kevin concluded his report by offering his impression of the interaction: "*He really wanted to help me out and seemed to be going to great lengths to find me a job.*"

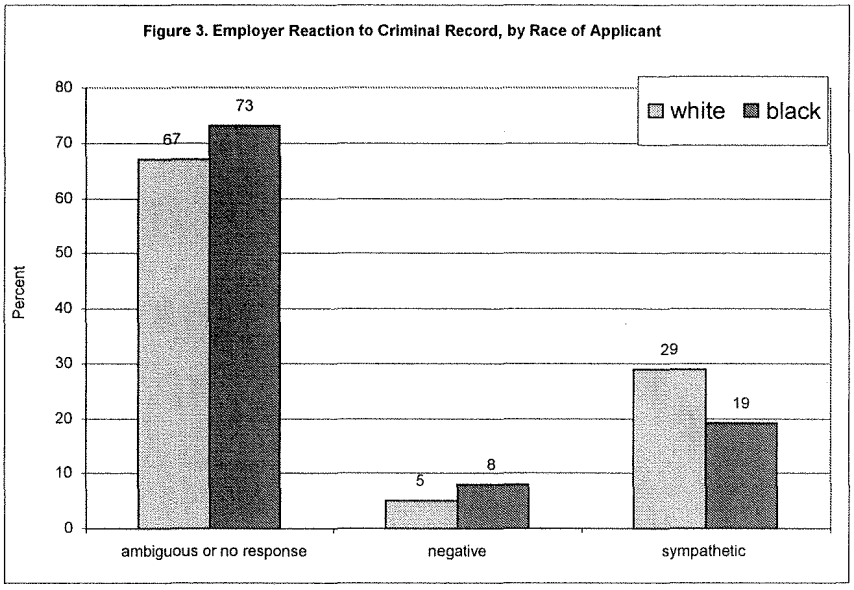
Once again, the ethnic solidarity expressed in this interaction appears to help establish a positive rapport between candidate and employer. Conversations with employers thus simultaneously offer the opportunity to present personalizing information about the applicant's work ethic and commitment to rehabilitation, but may also generate new bases for categorical distinctions.

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Employers thus appear to offer a range of reactions to ex-offender applicants, varying in terms of employers' comfort level in discussing criminal backgrounds and their evaluative assessments of this information. To examine employers' responses more systematically, we coded tester interactions with employers according to the nature of their response to the criminal record information, based on narrative data provided by the testers. Focusing on testers with criminal records who had personal contact with employers (roughly 50% of all tests), we code responses as: ambiguous or no response, negative response, and sympathetic response.

Looking to the results in Figure 3, we see that overall employers are most likely to avoid talking about the conviction altogether. Between two-thirds and three-quarters of employers either avoided the subject of the criminal record altogether or gave little indication of how they viewed the information. By contrast, less than 10 percent of employers made explicitly negative comments. Though the hiring outcomes from the audit study indicate a large negative impact of a criminal record on employers' evaluations, we see little of this reflected in their explicit comments to job applicants. If expressing a clear valence, employers were more likely to offer sympathetic reactions, with roughly 20 percent of employers coded as sympathetic toward the ex-offender applicant. We see some evidence that blacks were less likely to have a clear reaction from employers about their criminal background (73 vs 67%), more likely to receive a negative response (8% vs 5%) and less likely to receive a sympathetic response (19 vs 29%), though these differences are not statistically significant.

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.



Note: Includes those testers in the criminal record condition who interacted with the employer (n=144)

Overall, these results point to a reluctance among employers to address the criminal record issue head on, or to reveal their reaction to the record to the applicant in question. Our final question, then, considers the extent to which these differential responses in interaction correspond to differences in hiring outcomes.

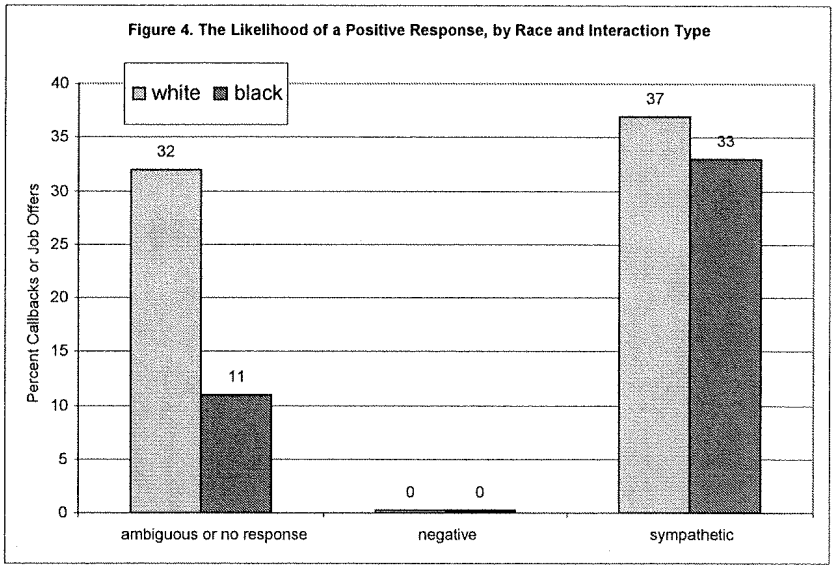
Relationship between Type of Interaction and Employment Outcomes

The nature of interaction between employer and applicant is significant only if it proves consequential for ultimate hiring outcomes. Matching interaction experiences with employment outcomes can provide some leverage on the pathways through which ex-offenders find opportunity. Figure 4 presents the percent of applicants with criminal records who received a callback or job offer, by race of the applicant and type of employer response. Not surprisingly, employment outcomes are most favorable among those who received a sympathetic response from employers. These employers were not simply paying lip service to the value of second chances, but demonstrated an actual willingness to hire ex-offenders. Among those who

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

received sympathetic responses from employers, whites were slightly more likely to receive an actual callback or job offer (37 vs 33%), though this difference is not statistically significant.

Showing even more consistency between interactions and outcomes, employers who expressed negative reactions to applicants with criminal records in no cases made offers or callbacks to these applicants.



Note: Includes those testers in the criminal record condition who interacted with the employer (n=144)

The group with less consistent results includes those employers who offered no reaction or ambiguous reactions to the criminal background. Among these employers, we see a large racial difference in employment outcomes with white applicants roughly three times more likely to receive a callback or job offer relative to blacks who had similar encounters (32 vs 11%). Note that outcomes for whites differed little for those who received ambiguous or no response and those who received a sympathetic response (32 vs 37%). Outcomes for blacks, by contrast, are quite different (11 vs 33%). Though it is not possible to directly interpret employers' underlying reactions, this evidence is consistent with the role of stereotypes inhibiting the acquisition and impact of individuating information. If employers who were concerned about the record among

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

black applicants chose to remain silent about the issue, the applicant would then have little opportunity to anticipate or address the employer's concerns. Where for white ex-offenders this reduced communication does not appear to have much of an impact, black ex-offenders seem to face significantly lower employment prospects as a result.

Conclusion to Stage One

The results of this study show a strong reluctance among employers to hire applicants with criminal records, and especially so when considering black ex-offenders. What is perhaps more noteworthy in these data relates to the cases in which testers with criminal backgrounds *were* given a chance at employment. Employment prospects improve significantly for applicants who have a chance to interact with the hiring manager, and more so among those who elicit sympathetic responses in the course of those interactions. To be sure, some of this variation is attributable to characteristics and preferences of the employers alone, with little or no effect of personal contact. Employers who are eager to hire will be more likely to meet with applicants on the spot, and those who are sympathetic toward ex-offenders will be more likely to express such sentiments in the course of interaction. At the same time, we suspect that the interaction itself plays a non-trivial role in this hiring process. Employers have many reasons to be concerned about applicants with prior histories of incarceration. Concerns about theft, violence, and drug use are each relevant, not to mention the more mundane concerns over worker reliability and performance. Personal contact with an applicant cannot reveal all of these issues, but can help to provide some signals as to the disposition of the applicant, and can help the employer to develop a "gut feeling" about whether or not this individual is likely to diverge from the stereotype of the ex-con.

Unfortunately, the ability to have such a hearing does not appear available to all applicants. Blacks were significantly less likely to be invited to interview by employers, offering

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

them fewer opportunities to present indicators at odds with their stigmatized group membership. Further, though the distribution of reactions from employers was roughly similar among black and white applicants with criminal backgrounds, actual employment outcomes differed for those who had little opportunity to discuss their criminal record: Among whites, these limited interactions did little to affect ultimate hiring outcomes, whereas for blacks positive responses appeared substantially reduced.

The importance of rapport-building in the employment process, particularly for applicants with stigmatizing characteristics, points to some important policy implications. Initiatives that facilitate the matching of workers with employers in ways that help to overcome these initial barriers may have a substantial impact. Intermediary organizations, for example, which develop relationships with employers willing to hiring ex-offenders and help them to train, screen, and place workers represents one important policy approach to bridging this divide.

Overall these findings suggest the importance of promoting social interaction with employers to encourage employment among workers with criminal records. Because of the large racial disparity in criminal convictions, job applicants with criminal background depend more on building rapport with prospective employers but are less likely to be given the opportunity because of their race. Policy intervention should aim to defuse stigma and provide employers with more information about their prospective workers. Job referral services that act as labor market intermediaries who vouch for job applicants might reduce the anonymous interaction of the low-wage labor market. Certificates of rehabilitation and public education campaigns might also weaken the effects of stereotyping.

As incarceration rates have increased over the last few decades, official criminality compounds the stigma of race and deepens of the economic disadvantage of young African American men. Instead of just adding to the deficits of low-skill black men, a criminal record

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

modifies the effect of racial discrimination, raising the bar to employment higher for blacks than similarly-situated whites. In this context, we can understand the growth of incarceration rates, and the racial disparities that characterize them, as producing an institutional racism with wide-reaching economic effects.

Stage Two: An Employer's Perspective

The audit study allows us to tell the story of the employment process from the job seeker's perspective. We can monitor the applicant's experiences at each stage of the hiring process, ultimately observing his 'success' or 'failure' in securing employment opportunities. By recording the differential outcomes of matched pairs of applicants, we can make inferences about employer preferences for particular worker characteristics; in particular, assessing the degree to which race and a criminal record represent important signals to employers in their evaluation of entry-level job candidates. And yet, we still know little about exactly how these signals work—what is the content of the signal that provides "useful" information to employers. A job applicant's criminal record may signal dangerousness, unreliability, or a variety of other unappealing characteristics yet we have little precise idea about employers' specific concerns. In order to unpack the meanings underlying employers' decisions, we turn our focus next to the employers themselves. By gathering insight into the structural and attitudinal factors that shape hiring outcomes, we can better understand the decision-making process that leads to the behavioral outcomes we observe. The dual perspectives afforded by this design—that of the job seeker and the employer—offer an innovative way to view the employment exchange.

In the second phase of this research, we conducted both a telephone survey and in-depth interviews with the same employers who participated in the audit study. All employers from the audit study sample were included for contact in the telephone survey, yielding a final sample of

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

385 employers (a response rate of 44%). On the survey, we included a wide range of questions regarding the employers' hiring policies, their preferences for a wide range of applicant types, and their experiences with recent hires. Of particular interest here, we included an extensive module which probed employers regarding their attitudes about and experiences with workers with criminal records.

Complementing the employer telephone surveys, we drew an additional subset from our employer sample (n=55) for the purpose of conducting in-depth in-person interviews. Through a series of open-ended questions, we solicited employers' own accounts of the jobs they seek to fill and the workers they find to fill them. During the interviews, which lasted between 45 minutes to 2 hours, employers were asked a series of questions probing perceptions of various groups of workers for an understanding of how race/ethnicity/criminal record may become important signals to employers making hiring decisions in the context of minimal information. These in-depth interviews allowed for more open-ended discussions of employers' concerns and considerations in hiring entry-level workers, providing greater insight into exactly how and why employers make the decisions they do.

In the following section, we provide an overview of employers' considerations and concerns regarding applicants with criminal backgrounds. We present the main results from the survey data paired with the rich qualitative data from our in-depth interviews to better understand employers' perspectives.

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Results[4]

In the survey we see a fairly strong reluctance to hire applicants with criminal records: Just over a quarter of employers (26%) indicated that they would probably or definitely be willing to hire an applicant with a criminal record, relative to 63% who said they probably or definitely would not, and only 11% who indicated that their decision would depend on the type or circumstances of the crime.[5] Several of the in-depth interviews further reflected employers' significant concerns about workers with criminal pasts. According to the manager of a retail store, *"I just think that people here would not allow someone with a criminal record to be around people that do not have a criminal record. It just is not safe. It is not fair to the other employees. To be around that kind of environment. You never know what can happen. If I had the power, I would make sure no one with a criminal record was ever hired here. It just is not right. I just think I wouldn't go for that."*

Despite the New York City Human Rights Law which prohibits blanket discrimination against individuals with criminal records, several employers candidly stated that any candidate with a criminal background would be immediately disqualified. According to one employer, "*It isn't worth the risk for us. That is company policy. We have a place on the application that asks about this. It is just not negotiable."* Even among employers from large national companies—whose attentiveness to legal requirements is typically far greater than small local employers—we see several cases of blanket bans on ex-offenders. Managers from three different national retail clothing stores told us of their company's policies against hiring ex-offenders. When asked how information about a criminal record might affect his evaluation of an applicant, the manager of

[4] In this report, we focus primarily on employers' views about applicants with criminal records. While race was also an important focus of this study as a whole, questions about racial attitudes and experiences are more difficult to probe using standard survey techniques. We did discuss these topics in more detail during the in-person interviews; further discussion of this analysis can be found in Pager & Karafin, 2009.

[5] Responses differed moderately by race of the employer, with black and Hispanic employers indicating a somewhat greater openness to hiring ex-offenders (25% and 30%, respectively) than white employers (23%).

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

one of these clothing chains replied, *"It doesn't impact me in any way, because at that point I'm not allowed to hire them."*

The manager of another retail chain elaborated on his understanding of his company's policy, and regulations about what information can be shared with applicants.

> *I: Okay, so I have a few more questions about your screening process. Let's say you have an applicant and you find out that they have a criminal record. How would you view this information?*
>
> *A: Well we can't reveal that information. I would just say to the applicant that at the moment it's not going to be possible because he can't qualify. You know, 'Thanks for applying but at this time it won't be possible. You are not right, we are looking for availability, it doesn't fit what we need, what the company is looking for.' You know, maybe you were making way too much money before, or something like that. I don't want to hurt anyone, you can't reveal that information, so you have got to keep it to yourself company-wise.*
>
> *I: So does the company have a formal policy about hiring ex-offenders?*
>
> *A: Yes.*
>
> *I: Does it matter based on the type of record, type of crime, and so forth?*
>
> *A: Anybody or anything or any sort of issues and we won't hire you.*

According to this manager, applicants are not meant to be told the specific reason for their dismissal when related to concerns about a criminal background, perhaps out of concern for legal challenges by the applicant in question. Nevertheless, the company appears to follow a clearcut rule against hiring ex-offenders.

The manager of a third national retail clothing store expressed similar concerns about legal challenges from applicants, though in this case inverts the law's intent, which is to promote individualized reviews rather than blanket restrictions. When asked whether exceptions could be made for exceptional candidates with criminal records this manager replied:

> *A: Absolutely not. If their background check does not come back clear, they would not be hired.*

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

I: I see. So that's just decided at the corporate level?

A: Correct…. From the company perspective, I think we're a large enough company now where you have to have some black and white rules, you know? You can't say, "Well, this is a drug conviction for possession. This guy's a really nice kid. We're going to hire him." Someone else comes back with stealing from another company. We're not going to hire him. That's opens you up for a lawsuit right there. And probably someone's never going to find out about that, because all their information's confidential, but it opens you up for a lawsuit, because you're not treating people fairly. So from the company's perspective, it makes sense to me to have certain rules that are just black and white. If your background check comes back negative, or it's not clear, sorry. Good luck on your search.

Whether based on formal company policy or the personal preferences of managers, that nearly two-thirds of employers expressed a reluctance to hire applicants with criminal records points the serious barriers facing young men leaving prison. Even as bipartisan support grows for supporting prisoner reentry (as evidenced by passage of the Second Chance Act), employers—who remain at the front lines of reentry efforts—remain highly skeptical of this population. Interestingly, employers' survey responses suggest an even stronger reluctance to hire ex-offenders than we see in the audit study, where roughly half the employers who hired at least one tester hired the tester presenting a criminal background. The disparity between survey and audit results may be do to a difference in attitudes versus behavior (e.g., even those who do not like the idea of hiring an ex-offender may do so if their labor needs are sufficiently great). We suspect, rather, that the primary difference results from our use of attractive, well-qualified testers who did not fit the employers' stereotype or expectation of an ex-offender. When answering the survey questions, employers are unlikely to have been imagining a candidate as appealing as those represented by our testers. If this is the case, the results of our audit study are likely to represent a highly conservative estimate of the effect of a criminal record on hiring outcomes, particularly as experienced by the more typical ex-offender who possesses lower levels of schooling and prior work experience.

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Employers' Concerns about Applicants with Criminal Records

When asked to describe their main concerns about hiring an applicant with a criminal record, employers focused on several key issues. A large number of employers focused on specific problem behaviors such as theft (21%), issues of violence and safety (9%), or the likelihood of reoffending (6%).[6] The manager of a record store explained, *"If it was theft and they were convicted I am going to stay away from that because it is very easy to steal from the warehouse and they have a lot of problems with people stealing CD's and DVD's. Those are tempting items. Now I know that is kind of discriminatory but I still do it. It is just one of those things that sits in the back of my head."* Likewise, the manager of a local wine store told us, "*We do have a big problem with theft, as you can imagine. So, I wouldn't want to hire someone who has had a problem in the past with stealing. I think this is a very tempting environment."* Concerns over theft came up often in our discussions with employers, particularly in retail environments or where employees are often handling cash or expensive merchandise.

A number of employers regarded violence and/or safety in the workplace as a primary concern when probed about hiring applicants with a criminal record. Many of these employers noted the discrepancy between non-violent and violent offenders as a critical element in their decision-making. Specifically, these employers expressed a fear of putting themselves, their employees, their customers, and the company as a whole, at risk. The manager of a testing company emphasized that he would avoid "*any kind of crimes that deal with bodily harm. Uh, you know, always concerned for the safety and welfare of my employees... I have to be worried about my employees."* Some also expressed concern that those with a violent past could not be

[6] The distribution of concerns differed somewhat across telephone survey and in-depth interview samples. In our in-person interviews with employers more than 40% regarded violence and/or safety in the workplace as a primary concern when probed about hiring applicants with a criminal record. Just over half (54%) regarded issues of theft a primary concern.

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

trusted with sensitive situations that could emerge – such a frustrating interaction with a customer, as in the case of this manager from a car dealership:

> *I mean if you have somebody with armed robbery, or rape, or something like that, this is a violent person and I don't want that here. This is a very sensitive business. People come in and they expect a lot. We don't have a good reputation to begin with. People would prefer to go to the dentist rather than buy a car. (laughs) People come in sometimes and they will be short-tempered. Oh my god. They will say something like, oh my car is scratched, and they will come in screaming and yelling. You have to be able to diffuse that situation. I need somebody with patience. I need somebody that can talk to the person, calm them down, and rectify the problem. What I don't need is someone that is just going to turn around to someone that is screaming at them and punch them in the face.*

In addition to concerns over problem behaviors, employers also discussed character flaws reflected by a criminal history, such as concerns over honestly/trustworthiness (16%) or character/reliability (3%). When asked about her primary concerns about considering an applicant with a criminal record, one retail employer stated simply, "*Well, mostly the concern of trust. Can the person be trusted? This would be my biggest concern.*" Another employer went further in explaining his view of an individual with a criminal record:

> *I would just be reluctant, you know, about morality, their character. Like a pedophile or someone who has a record of that type. Even though we don't have kids in our place I would still be looking at that person, not fearful, but with contempt. So maybe I wouldn't hire that person. Maybe I am being prejudiced but I think it is a legitimate concern. I would really really have to give it some thought before I would hire someone who was a criminal in the past. The person has to have references a mile high for me to feel confidant. Somebody would have to call me and say I know the person for so long and he is under my care, he is in a program right now.*

Finally, a few employers cited specific hassles associated with hiring ex-offenders as a primary reason for avoiding such applicants. According to a tour company manager:

> *I don't want to have to deal with parole officers, probation officers, sending letters, having somebody come by to check up on them, and that kind of stuff. I can't have that type of distraction. I am just too busy for that. You know, as good as somebody might be, I have 40 other employees to think about and their welfare and their well-being. I don't know what kind of people the ex-offender may be associated with that might be coming here to look for them. It is kind of like, this is a business. We are publicly traded on the New York Stock Exchange.*

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

An employer from a video rental store emphasized a similar kind of hassle, though in this case related more to his practice of paying workers off the books:

> *I mean the only thing [that would make me reluctant to hire someone with a criminal record] would be that I do hire a lot of people off the books. And ex-offenders do have to show a pay stub to their parole officer, so that could be an issue for me. I would have to work it out where they were on the books I guess.*

Though many employers expressed strong opinions about how they would view an applicant with a criminal record, a sizeable fraction (31%) indicated that their evaluation of the candidate would depend on the severity, nature, and circumstances of offense.[7] According to a toy store manager, *"It depends kind of on the situation. How long ago was it, what was it? There are people that have turned around and they are fine now. I am sure that some of them have worked for us and we don't even know. It is hard to know."* The manager of an employment agency echoed these sentiments, emphasizing the severity of the crime as a key determinant of his evaluation: "*Well, it depends on what the crime is, and how long they were in prison. If they were in for 20 years, that would be bad. You know, you wouldn't be in jail for 20 years for stealing a bagel from a restaurant. Length of time is important. That can show if it was a very serious thing.*"

When we asked the manager of a bakery about her views on hiring an applicant with a criminal record, her initial response was very negative, but she then conditioned this reaction with attention to the specifics of the offense:

> *I: If you found out an applicant had a criminal record, how would you view this information?*

[7] This distribution is based on an open-ended survey that asked, "What would be your main concern about hiring an applicant with a criminal record?" Earlier in the survey, we asked a more general close-ended ended question which asked, "Would you be willing to hire an applicant with a criminal record?" In response to this question (discussed above) nearly 90 percent offered a firm positive or negative response, with only 11 percent indicating that their decision would "depend on the crime." This latter category represents a spontaneous response—not one solicited by the close-ended format—suggesting that these employers were resistant to being placed in a category they felt inappropriately reflected their views. When allowed to answer in the open-ended format, by contrast, far more conveyed variability in their responses. This finding is consistent with other surveys (e.g., Holzer et al., 2002, Pager, 2007), who report that up to 30 percent of employers report that their response "depends on the crime."

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

> *A: Eww. (pauses) I'm not sure. I don't like criminals (laughs). But, you know, I guess it depends. What the crime was, how long ago it was, how rehabilitated they seem. I wouldn't make a definite generalization and say no I wouldn't hire a criminal, or yes, it wouldn't matter at all. You know, it depends on the individual and the circumstance.*

Another employer from a dessert chain emphasized the importance of offense severity, as well as evidence of rehabilitation and future orientation.

> *Well, I would first look at what the offense was, if it was a misdemeanor or a felony. From there I would speak to the person, and try to get a feel for what they were trying to do. What is their plan long-term, and what are their goals? I definitely wouldn't have any problem hiring someone with a prior conviction, though I don't get to make the final decision here. For someone who I feel would work out great, I would certainly put my neck on the line to help them get hired here. I don't think it is the deciding factor though, whether or not they have a record. I think their character is what matters. I do think that the other managers and the owners would view a criminal record as a red flag, since you did have an incident. They would be very reluctant to hire someone like that. But, for me, I would really hope that they would at least consider them.*

The manager of an ice cream chain emphasized the distinction between violent and non-violent offense types:

> *Last year, this kid came to me because he was kicked out of his high school because he had a drug problem. He said that he went through rehab and all this stuff and is trying to turn his life around. I like that, so I gave him a job. He worked out great. Or, you know, if someone was convicted of something that I felt was a crime. Marijuana possession or something like that. I really wouldn't care too much about. Anything violent, if I was privy to that knowledge, I would definitely not hire them. Violence isn't good. I wouldn't want that vibe around here. Especially if I knew the person had a history of it.*

In each of these cases, employers appear sensitized to the diversity of criminal records, and apply varying degrees of scrutiny depending on how they evaluate information about risk. Though employers emphasized a wide range of distinctions, the most common appeared to be distinctions according to crime type (with violent offenses and theft eliciting the greatest concern) and signs of rehabilitation (e.g., time since offense, evidence of drug treatment, indications of sincere repentance, etc.).

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Legal Liability

A growing literature emphasizes employers' concerns over legal liability as a key factor reducing the job prospects of ex-offenders. Indeed, a few employers in our sample did raise such concerns. For example, an employer for a large national company explained: *"There's a legal category called negligent hiring, if we knowingly hired somebody that had a criminal record, say someone was charged with physical assault, we would subject to a lawsuit under negligent hiring we should have known this person was apt to act out violently in the workplace."* The manager of a courier company, discussing his reluctance to hire anyone with a history of violent crime, touched on similar themes, though couched as an insurance concern: "*It's an insurance problem. I can't, I can't get insurance coverage, I can't bond, and I'm going to be sitting here wondering, if he has a problem at work, is he going to smack somebody, one of my customers, or do something like that in the street?"* For these employers, workers with criminal records pose risks that extend beyond specific acts of violence or theft to represent significant financial liability.

Interestingly, however, less than one percent of the employers in our telephone survey sample mentioned the issue of legal liability as a key concern. Further, when asked about the issue directly in the in-person interviews, most employers had never heard of negligent hiring laws. Even those who were familiar with these laws appeared more concerned over the behaviors themselves (violence, theft) than potential legal liability resulting from those behaviors. According to a retail store manager,

> *I mean, it is not in the forefront of my thoughts. I think people with a violent nature, and anger problems, seem to have those regularly. They are explosive. I try my best to be in a situation where I don't have to deal with that. Dealing with the public can be a trying experience.... But in general, I am not usually thinking, oh, I don't want to take this person because I am afraid I am going to have some legal issue.*

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

The manager of a record distribution company dismissed the notion that individuals with criminal records pose any particular liability:

> *I: Are there ever any legal issues that you would worry about in hiring ex-offenders?*
>
> *A: Umm actually no. To tell you the truth no. You know what, I honestly have never thought about this before. (laughs) But thinking about it now, if it were to happen, I think there would be some type of liability case whether or not the person had a prior record. Yeah.*

The owner of a supply company felt confident that the ex-offenders referred to his business were insured against possible liability:

> *I: Are there ever any legal issues that you would worry about in hiring ex-offenders?*
>
> *A: None. If I am hiring one of those it is from one of those programs and they are guaranteed and they are bonded. If they do anything here then I would be reimbursed.*

These findings stand against the growing literature emphasizing negligent hiring laws as a major reason that employers are reluctant to hire ex-offenders (e.g., Bushway et al., 2007). It is likely that employers for large national companies (underrepresented by our sample) have more elaborate legal regulations related to hiring,[8] norms about which may ultimately trickle down to smaller employers over time. Nevertheless, our results suggest that negligent hiring laws themselves are not at present a major barrier to employment for ex-offenders. Rather, possible problem behaviors themselves (e.g., violence, theft), rather than legal liability for those behaviors, appear to drive employers concerns about this population.

In general, laws appear to have only marginal effects on employer attitudes about hiring ex-offenders. As discussed earlier, New York City offers protection from blanket discrimination on the basis of a criminal record, though many employers are unaware of the specific legal

[8] This is also consistent with our finding from the audit study that large employers are less likely to hire applicants with criminal records (opposite to the pattern of hiring racial minorities).

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

provisions and/or seem unconcerned about their effects. Only a few employers explicitly referred to antidiscrimination law in discussing the evaluation of applicants with criminal records. For example, one manager of a national clothing chain explained:

> *Well legally you don't turn an application away as long as they reveal that on their application and they are forthcoming. You know it really depends. I guess it depends on what they did. I would always pass it to HR if it was someone that I thought was worthy of the job, I would always partner with Human Resources first. You don't want to discriminate. It is against the law to discriminate like that. And we have had that, where we hired those kind of people, and that is what I do, I partner with our HR department. They help to make sure that I am not breaking any rules, you know.*

In general, however, special legal protections appeared to have little effect on employer's attitudes about hiring ex-offenders. Even among employers who were aware of antidiscrimination laws protecting people with criminal records (a question we explicitly asked on the survey), fully 60% maintained that they would not be willing to hire an ex-offender. While lower than the 70% unwilling to hire ex-offenders among those who do not believe ex-offenders are covered by antidiscrimination law, in either case a large majority of employers remain unwilling to hire applicants with criminal backgrounds.

In contrast to sanctions, we see some evidence that incentives are associated with changes in employer behavior. Tax credits and the federal bonding program, in particular, appear to play some role in employers' considerations. Of those who had hired an ex-offender in the past year, roughly 12% had claimed the Work Opportunity Tax Credit (WOTC) and 4% have had their ex-offender employees covered by the Federal Bonding Program. Of those who had claimed the WOTC, fully 78% indicated they would be willing to hire ex-offenders in the future (7 of the 9 employers who had claimed the WOTC), relative to only 43% of those who had hired ex-offenders in the past but did not claim tax credit for those hires. Both employers who had bonded ex-offender employees through this program indicated they would be willing to hire ex-offenders in the future, though the sample size here is too small to generate reliable comparisons.

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Though only a small minority of employers who hire ex-offenders make use of tax credits and federal bonding, these incentives do appear to increase employers' willingness to continue hiring ex-offenders in the future.

An Uncomfortable Subject

The results of the audit study suggested that, though employers clearly notice criminal background information in reviewing job applicants, many are reluctant to discuss the subject explicitly. Likewise, in our interviews with employers, some emphasized the discomfort and anxiety that accompanies evaluations of applicants with criminal pasts. In some cases, employers emphasized not just interpersonal anxiety but also the perception that they were legally constrained from probing applicants about private legal matters. For example, one employer expressed the belief that criminal record information was off-limits in screening applicants:

> *A: In all honesty I am not really supposed to know what they have done anyway so most of the time I really don't know unless they voluntarily tell me. Usually I just don't know.*
>
> *I: You don't ask this on the application?*
>
> *A: No, you can't ask them. It is against the law.*

Likewise, the manager of a local restaurant expressed the belief that employers are prohibited from asking applicants about their criminal pasts:

> *We do the criminal check. We can't ask them, 'Are you a criminal or not?' That's why we check, the criminal check. So it says everything. Also it's against the law, you can't ask them, 'Are you criminal? Are you?'*

For the manager of a national restaurant chain, perceived limits on acquiring information about a criminal record leads to a better-safe-than-sorry policy of blanket exclusion:

> *This person is probably not trustworthy. They're not dependent, and if you're using drugs, how much are you using drugs. And I cannot go and ask you a question. 'All right, you [were] convicted of using drugs, how much drugs [did] you use?' I cannot ask that question. My question solely has to be based on your*

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

> *quality and your ability to do the job. So, being that I can't ask that question straightforward I have to just use my own judgment and say I should probably stay away from this person. Just to be safe.*

In this case, the employer's misperception of the kinds of questions that are lawfully appropriate in job interviews leads him to refrain from probing about the context of the conviction and instead avoid ex-offenders altogether. As we saw above in the interactions between applicants and employers in the audit study, employers who refrained from discussing the record with job applicants were much less likely to hire these candidates, particularly in the case of black ex-offenders.

The additional discomfort of race was raised explicitly by an employer for a tour company, who commented on the problem of being perceived as racist for inquiring about an applicant's criminal background: *"I try not to ask [about a criminal background in the interview] because sometimes people are like, 'What, are you racially profiling me?' So I do ask everybody but it is very hard to ask. It is not an easy question...."* This employer appears to struggle with concerns about how to talk about an applicant's criminal history without appearing to be unfairly targeting black applicants for scrutiny. Given the charged issues around race and perceived discrimination, employers may be especially reluctant to voice concerns about criminal records with black applicants, perhaps feeding in to some of the racial disparities reported above. If employers feel uncomfortable inquiring about an applicant's criminal background and gathering additional information about the context of the conviction, they may instead choose to simply rule out these applicants altogether.

Prior Experiences with Ex-Offenders

According to the survey, about 30 percent of New York City employers have hired an applicant with a criminal record over the past year. In both the survey and during in-person interviews, we asked employers to tell us about their previous experiences with workers with criminal

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

backgrounds. In a few cases, employers reported negative experiences with ex-offender employees. For example, the manager of a dessert chain told us about one recent experience: *"Well, I have had an applicant who worked here, who was in a juvenile hall. She had a history with drugs. And, umm. She was trying to clean up her act. Eventually I had to fire her. She was showing up late, going out all night."* Though this employer remained open to the possibility of hiring ex-offenders in the future, the experience points to some of the potential problems posed by individuals with histories of substance abuse. For others, negative experiences left them extremely reluctant to hire applicants with criminal backgrounds in the future. For example, the manager of a retail store explained, *"We recently had a receptionist who was stealing from the register. We had known about a drug history past. And she was in recovery and so we gave her a chance. I don't think that my boss would take that chance again."* In this case, one negative experience can lead to strong generalized views about the risks posed by ex-offenders.

Apart from a small number of strongly negative experiences, for the most part employers reported mixed or positive experiences with ex-offenders, leaving them open to hiring additional ex-offenders in the future. According to the manager of a clothing chain, *"So, in terms of the criminal record, it does not mean that I wouldn't hire them. I have hired them before. I have had two great employees that had criminal records, and I have had one horrible one. So it is kind of hit or miss."* Similarly, the manager of a courier company explained, *"Very rarely do I have guys from the facility or programs that I work with that end up getting in trouble here. In six years of hiring I would say maybe two of the ex-offenders I have hired have done something on the job that they weren't supposed to do."*

Consistent with these more positive experiences, in general we find evidence that previous experience with ex-offenders increases the likelihood that an employer would be

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

willing to hire ex-offenders in the future. Of those who had hired an ex-offender in the past year, fully 43% indicated they would be willing to hire an applicant with a criminal record in the future, relative to only 17% of those without recent experience hiring ex-offenders. Some of this disparity is likely to simply reflect the diverse applicant pools different employers draw from; but we also see some evidence that the experience of working with ex-offenders actively reduces employers' concerns. Indeed, roughly 85% of employers who have hired ex-offenders in the past year report their work performance to be at or above average, suggesting that in general employers are having fairly positive experiences with this group of employees. The following excerpt comes from an interview with the manager of a national retail store:

> *I: How have the employees you have hired with a record worked out compared to employees you have hired without a record?*
>
> *A: Umm, actually they have pretty much been good employees. I don't know if it is because they had to keep the job. (laughs) But you know I didn't have a problem with them. I mean a few people that have applied with a criminal background have been turned away, but it was because like one told me he had been convicted of attempted murder. So that kind of scared me a little. I went through HR and it was time consuming but we decided not to hire him. But that was a good thing. I have hired people though and I haven't really had bad experiences with these people.*

Several employers emphasized particular advantages of ex-offenders as employees, in that they may be more likely to take the job seriously and value the opportunity more than workers with wider options. An employer for a manufacturing company described his experiences:

> *I: Is there any way to predict whether or not an applicant will be likely to steal, aside from them having a criminal record?*
>
> *A: You know you really can't tell. The majority of them are fine. These are just examples that I'm pulling out because you are asking me. So those two are the only ones. We have another fellow downstairs who is here every day and is very good. We had to sign for him for being in the halfway house and he is a wonderful guy. A lot of people here need their jobs so they are very committed.*
>
> *I: How many ex-offenders do you currently have on staff?*
>
> *A: Let's see. We have about 4 right now and really they are working out fine.*

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

The manager of a specialty foods store reports similar experiences, emphasizing how committed ex-offenders are to staying out of prison:

> *A: We have actually hired people on work-release programs. So that is not a problem. Having a problem, giving someone a second chance is not a problem. It is when they are not reformed and they continue. And honestly, my feeling on that is that you are so close to getting out that you are not going to screw up. You have a taste of freedom but you have to go back at a certain time after work. You have an hour and a half or whatever it is that you have to get home. So you are too close to tasting that freedom that you are not going to screw up here. So that is my theory on why they are going to be good for us.*
>
> *I: How did the ex-offender employees you hired work out?*
>
> *A: Oh they worked out good. I think that system really works.*

An employer for a courier company echoed these sentiments, "*I've had a couple of people that were upfront, said they had criminal backgrounds, and I checked in with their parole or probation officers, and they worked out pretty well.... Matter of fact they, they turned out to be some of the better workers. 'Cause they're trying to make another shot at life for themselves.*"

Overall, then, employers reported generally positive experiences with ex-offenders and remain more open to hiring ex-offenders in the future relative to those with no recent experience. Further, employers who have had recent experience working with ex-offenders report fewer serious concerns about this group relative to employers who have no recent experience. As shown in Table 1, fully a quarter of employers with no recent experience with ex-offender employees are concerned about theft by this group, relative to less than half that proportion of employers who have hired ex-offenders in the past year. Likewise, 18% of employers unfamiliar with ex-offenders are concerned about honesty or trustworthiness, relative to only 11% of those who have recently hired ex-offenders. Employers who have worked with ex-offenders in the recent past are more likely to acknowledge wide variation among the ex-offender population, with 45% indicating that their willingness to hire an ex-offender would depend on the severity, nature, or circumstances of the crime (relative to just 26% of those employers less experienced

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

with this population). Interestingly, those with recent experience with ex-offender employees were more than twice as likely to report concerns over reliability (4.8% vs 2.2%) and drug/alcohol use (2.4% vs 0.44%), though each of these concerns came up among less than five percent of employers overall.

Table 1. Primary Concern about Hiring an Applicant with a Criminal Record, by Prior Experience with Ex-Offender Employees

	Hired Ex-Offender in Past Year?		
Concern	no	yes	total
depends on severity/circumstance	26.22	44.48	31.17
theft/stealing	24.89	12.05	21.43
honesty/trustworthiness	17.78	10.84	15.91
violence/safety	8.89	9.64	9.09
concern over worker quality	3.56	4.82	3.90
will re-offend	6.67	4.82	6.17
reliability	2.22	4.82	2.92
drugs/alcohol	0.44	2.41	0.97
constraints due to parole	1.33	1.20	1.30
legal/liability concerns	0.89	2.41	1.30
other	7.11	2.41	5.84
total	100	100	100

In general, then, experience with ex-offender employees appears to reduce employers' major anxieties about this group (theft, honesty), while increasing recognition of the diversity of the ex-offender population and the need to carefully review the context and severity of the individual's criminal history.

Mirroring findings about the importance of personal contact presented above, these results suggest that the more familiarity employers have with ex-offenders—either based on previous employees or through personalizing information about a particular ex-offender applicant—the more open an employer is likely to be with respect to hiring such candidates.

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Conclusion to Stage II

Many studies of employers focus merely on employers' own accounts, taking at face value the concerns, priorities, and practices they describe (e.g., Holzer, 1996; Waldinger and Lichter, 2003). This research, by contrast, is solidly grounded in external observation of employer behavior. The audit study demonstrates the selection process in real employment interactions; the interviews with employers then help us to understand the thought processes and considerations that may have led to the observed outcomes. In this way, the research offers the unique opportunity to examine what employers say in the context of what they actually do.

Substantial investment has gone into developing programs that facilitate offender reentry. Such efforts routinely emphasize the fundamental importance of employment, typically with a focus on improving the skills and resources of the offenders themselves (Travis, 2003; Petersilia, 2003; SVORI, 2004). We applaud these efforts, and hope to build on their insights. Developing an approach complementary to offender-focused strategies, we place emphasis on the demand side of the labor market. Improving the job readiness of ex-offenders can only be effective if employers are willing to hire individuals with criminal records. Understanding the strength and sources of employers' aversion to ex-offenders can direct us in shaping strategies to facilitate effective job placement for this group.

This research raises important questions about the kinds of policies that may make employers more likely to hire ex-offenders. In particular, we believe this research has important implications for the work of parole officers and employment mediators. Given the findings that employers' anxieties about hiring ex-offenders focus more on risk (of violence/theft) rather than skills, these groups can play an invaluable role in helping to reduce the uncertainty associated with hiring individuals with criminal pasts. Parole officers and employment mediators can speak knowledgeably to the concerns of employers, helping them to better understand the risks

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

associated with different kinds of offenders, as well as in providing additional sources of oversight for these workers. In the spirit of Petersilia's (2003) suggestion of establishing community partnerships, this research supports the development of relationships between parole agencies, intermediaries, and employers, improving the prospects for reintegration and public safety overall.

Dissemination Strategy

We have presented the results of this research in a number of public and academic forums. Early versions of the audit research were presented at the annual meeting of the American Sociological Association, the Population Association of America, and at numerous academic departments. In addition, we were invited to present our results for a regional meeting of the EEOC in New York City, for the New York City Council Civil Rights Committee, and at the New York State Attorney General's Office on Civil Rights. We were also invited to testify at a hearing of the EEOC in Washington DC concerning employment regulations governing the use of criminal records. Our research was featured in a conference sponsored by the New York City Commission on Human Rights, with hundreds of New York employers in attendance; and at a public forum sponsored by the John Jay College of Criminal Justice (leading to coverage in the *New York Times* and several local media outlets). Finally, we are publishing our results in a number of academic journals, including forthcoming articles in the *American Sociological Review*, the *Annals of the American Society of Political and Social Sciences*, and the *University of Chicago Legal Forum*.

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

References

Ayres, Ian and Peter Siegelman. 1995. "Race and Gender Discrimination in Bargaining for a New Car." *The American Economic Review* 85(3):304-321.

Bartik, Timothy. 2001. *Jobs for the Poor: Can Labor Demand Policies Help?* New York: Russell Sage Foundation.

Becker, Howard. 1963. Outsiders: Studies in the Sociology of Deviance. London: Free Press

Bendick, Marc, Jr., Charles Jackson, and Victor Reinoso. 1994. "Measuring Employment Discrimination through Controlled Experiments." *Review of Black Political Economy* 23: 25-48.

Bonczar, Thomas P. and Allen J. Beck. *Lifetime Likelihood of Going to State or Federal Prison.* Washington, DC: U.S. Department of Justice.

Borjas, George J. 1999. Heaven's Door: Immigration Policy and the American Economy. Princeton, NJ: Princeton University Press.

Bushway, Shawn D. 1997. "Labor Market Effects of Permitting Employer Access to Criminal History Records." Working Paper. University of Maryland, Department of Criminology.

Bushway, Shawn D., Michael A. Stoll, David F. Weiman. 2007. *Barriers to Reentry?: The Labor Market for Released Prisoners in Post-Industrial America.* New York: Russell Sage Foundation.

Bradburn, N.M. 1983. "Response Effects." In Rossi, P., J. Wright, and A. Anderson (Eds.). *Handbook of Survey Research.* New York: Academic Press, pp.289-328.

Braman, Donald. 2004. *Doing Time on the Outside: Incarceration and Family Life in Urban America.* Ann Arbor: University of Michigan Press.

Branham, Lynn S. 2002. *The Law of Sentencing, Corrections, and Prisoners Rights.* 6th Edition. St. Paul: Minnesota: West Group.

Bureau of Justice Statistics Bulletin. 2001. *Prisoners in 2000,* by Allen J. Beck and Paige M. Harrison. August, Washington, D.C. NCJ 188207.

Bureau of Justice Statistics, Special Report. 2003. *Education and Correctional Populations.* By Caroline Wolf Harlow. January, Washington, D.C., NCJ 195670.

Bussey, Jenny and John Trasvina. 2003. "Racial Preferences: The Treatment of White and African American Job Applicants by Temporary Employment Agencies in California." Working Paper, Discrimination Research Center, The Impact Fund.

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Cancio, A. Silvia, T. David Evans, and David J. Maume. 1996. "Reconsidering the Declining Significance of Race: Racial Differences in Early Career Wages." *American Sociological Review* 61:541-56.

Crandall, C.S. 1994. "Prejudice against Fat People: Ideology and Self-Interest." *Journal of Personality and Social Psychology* 66: 882-894.

Cross, Harry, Genevieve Kenney, Jane Mell, and Wendy Zimmerman. 1989. *Differential Treatment of Hispanic and Anglo Job Seekers: Hiring Practices in Two Cities.* Washington, DC: Urban Institute Press.

Cross, Harry, et al. 1990. *Employer Hiring Practices: Differential Treatment of Latino and Anglo Job Seekers.* Washington, DC: Urban Institute Press.

Culp, Jerome and Bruce Dunson. 1986. "Brothers of a Different Color: A Preliminary Look at Employer Treatment of White and Black Youth." In Freeman, Richard B. and Harry J. Holzer (eds). 1986. *The Black Youth Employment Crisis.* Chicago, IL: U Chicago Press for National Bureau of Economic Research.

D'Souza, Dinesh. 1995. *The End of Racism: Principles for a Multiracial Society.* New York: Free Press.

Duster, Troy. 1997. "Pattern, Purpose and Race in the Drug War: the Crisis of Credibility in Criminal Justice." In Crack in America: Demon Drugs and Social Justice, edited by Craig Reinarman and Harry G. Levine. Berkeley, CA: University of California Press.

Espino, Rodolfo and Michael M. Franz. "Latino Phenotypic Discrimination Revisted: The Impact of Skin Color on Occupational Status." *Social Science Quarterly* 83(2): 612-623.

Farkas, George. and K. Vicknair. 1996. "Appropriate Tests of Racial Wage Discrimination Require Controls for Cognitive Skill: Comment on Cancio, Evans, and Maume." *American Sociological Review* 61:557-60.

Fix, Michael and Raymond J. Struyk (Eds.). 1993. *Clear and Convincing Evidence: Measurement of Discrimination in America.* Washington, DC: Urban Institute Press.

Freeman, Richard B. 1995. "The Labor Market." Pp. 171-192 in *Crime*, edited James Q. Wilson and Joan Petersilia.San Francisco, CA: Institute for Contemporary Studies Press.

Freeman, Richard B. and Harry J. Holzer (eds). 1986. *The Black Youth Employment Crisis.* Chicago, IL: U Chicago Press for National Bureau of Economic Research.

Hakken, Jon. 1979. *Discrimination against Chicanos in the Dallas Rental Housing Market: An Experimental Extension of the Housing Market Practices Survey.* Washington, DC: U.S. Department of Housing and Urban Development.

Harris, Anthony. 1976. "Race, Commitment to Deviance and Spoiled Identity." *American Sociological Review* 41:432-42.

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Heckman, James and Peter Siegelman. 1993. "The Urban Institute Audit Studies: Their Methods and Findings: Response to Comments by John Yinger" Pp. 271-75 in *Clear and convincing evidence: Measurement of Discrimination in America*, edited by Michael Fix and Raymond Struyk. Lanham, MD: University Press of America.

Hogg, Robert V. and Allen T. Craig. 1989. *Introduction to Mathematical Statistics*. Fourth Edition. New York: Macmillan.

Holzer, Harry J. 1996. *What Employers Want: Job Prospects of Less-Educated Workers*. New York: Russell Sage Foundation.

Holzer, Harry, Steven Raphael, and Michael Stoll. 2001. "Perceived Criminality, Criminal Background Checks and the Racial Hiring Practices of Employers." *Working Paper*.

Holzer, Harry, Steven Raphael, and Michael Stoll. 2002. "Will Employers Hire Ex-offenders? Employer Preferences, Background Checks, and Their Determinants." Institute for Research on Poverty Discussion Paper 1243-02.

Kirschenman, Joleen and Katherine Neckerman. 1991. "We'd Love to Hire Them, but....": The Meaning of Race for Employers," in Christopher Jencks and P. Peterson, *The Urban Underclass*, Washington, D.C.: Brookings Institute.

Kluegel, James R. and Lawrence Bobo. 2001. "Perceived Group discrimination and Policy Attitudes: The Sources and Consequences of the Race and Gender Gaps." In Urban Inequality: Evidence from four Cities, edited by A. O'Connor, C. Tilly, and L.D. Bobo. New York: Russell Sage Foundation.

Loury, Glenn C. 1977. "A Dynamic Theory of Racial Income Differences." In Wallace, PA and La Mond, AM (Eds). *Women, Minorities, and Employment Discrimination*. Lexington, MA: Heath, pp.153-86.

Milgram, Stanley. 1974. *Obedience to Authority: An Experimental View*. New York: Harper and Row.

Moss, Philip, and Chris Tilly. 1996. "'Soft Sills' and Race: An Investigation of Black Men's Employment Problems." *Work and Occupations* 23(3): 256-276.

Nurse, Anne. 2002. *Fatherhood Arrested: Parenting from Within the Juvenile Justice System*. Nashville, TN: Vanderbilt University Press.

Neal, Derek, and William Johnson. 1996. "The Role of Premarket Factors in Black-White Wage Differences." *Journal of Political Economy,* 104(5):869-895.

Neumark, David. 1996. "Sex Discrimination in Restaurant Hiring: An Audit Study." *Quarterly Journal of Economics* XX: 915-40.

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Pager, Devah, Bruce Western, and Bart Bonikowski. 2009. "Race at Work: Results from a Field Experiment of Discrimination in a Low Wage Labor Market." *American Sociological Review*, forthcoming.

Pager, Devah and Diana Karafin. 2009. "Bayesian Bigot? Statistical Discrimination, Stereotypes, and Employer Decision-Making." *Annals of the American Academy of Political and Social Sciences* 621 (January):70-93.

Pager, Devah. 2007. *Marked: Race, Crime, and Finding Work in an Era of Mass Incarceration.* Chicago: University of Chicago Press.

Pager, Devah. 2003. "The Mark of a Criminal Record." *American Journal of Sociology* 108(5): 937-975.

Petersilia, Joan. 2003. *When Prisoners Come Home: Parole and Prisoner Reentry.* Chicago, IL: University of Chicago Press.

Pettit, Becky and Bruce Western. 2001. "Inequality in Lifetime Risks of Incarceration." Paper presented at the annual meetings of the Population Association of America, Washington DC.

Portes, Alejandro and Cynthia Truelove. 1987. "Making Sense of Diversity: Recent Research on Hispanic Minorities in the United States." *Annual Review of Sociology* 13: 359-85.

Steele, Shelby. 1991. *The Content of Our Character: A New Vision of Race in America.* New York: Harper Perennial.

Sullivan, Mercer. 1989. *"Getting Paid:" Youth Crime and Work in the Inner City.* Ithaca, NY: Cornell University Press.

SVORI, 2004. *National Portrait of SVORI: Serious and Violent Offender Reentry Initiative.* http://www.svori-evaluation.org/SVORINationalPortrait.pdf

Travis, Jeremy. 2003. *Prisoners Once Removed.* Washington, DC: Urban Institute Press.

Travis, Jeremy. Forthcoming. *But They All Come Back.* Washington DC: Urban Institute Press.

Turner, Margery, Michael Fix, and Raymond Struyk. 1991. *Opportunities Denied, Opportunities Diminished: Racial Discrimination in Hiring.* Washington, DC: Urban Institute Press.

U.S. Census Bureau. 2003. "Hispanic or Latino by Type: 2000; Geographic Area: New York County, New York." Census 2000 Summary File 1. American FactFinder.

Wacquant, Loic. 2000. "The New `Peculiar Institution:' On the Prison as Surrogate Ghetto." *Theoretical Criminology* 4:377-89.

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Waldinger, Roger D. and Michael I Lichter. 2003. *How the Other Half Works: Immigration and the Social Organization of Labor*. Berkeley, CA: University of California Press.

Western, Bruce. 2002. "The Impact of Incarceration on Wage Mobility and Inequality." *American Sociological Review* 67:477-98.

Western, Bruce and Katherine Beckett. 1999. "How Unregulated is the U.S. Labor Market? The Penal System as a Labor Market Institution." *American Journal of Sociology* 104(4):1030-60.

Western, Bruce and Becky Pettit. 2000. "Incarceration and Racial Inequality in Men's Employment." *Industrial and Labor Relations Review* 54(1): 3-17.

Western, Bruce, Becky Petit, and Josh Guetzkow. 2002. "Black Economic Progress in the Era of Mass Imprisonment." Pp. 165-80 in *Invisible Punishment: The Collateral Consequences of Mass Imprisonment*, edited by Meda Chesney-Lind and Marc Mauer. New York: New Press.

Western, Bruce, Jeffrey R. Kling, and David F. Weiman. 2001. "The Labor Market Consequences of Incarceration." *Crime and Delinquency* 47:410-27.

Wienk, Ronald E., Clifford E. Reid, John C. Simonson, and Frederick J. Eggers. 1979. *Measuring Discrimination in American Housing Markets: The Housing Market Practices Survey.* Washington, DC: U.S. Department of Housing and Urban Development.

Wilson, William Julius. 1987. *The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy.* Chicago: The University of Chicago Press.

Wilson, William Julius. 1996. *When Work Disappears: The World of the New Urban Poor.* New York: Vintage Books.

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Sequencing Disadvantage: Barriers to Employment Facing Young Black and White Men with Criminal Records

By
DEVAH PAGER,
BRUCE WESTERN,
and
NAOMI SUGIE

In this article, the authors report the results of a large-scale field experiment conducted in New York City investigating the effects of race and a prison record on employment. Teams of black and white men were matched and sent to apply for low-wage jobs throughout the city, presenting equivalent resumés and differing only in their race and criminal background. The authors find a significant negative effect of a criminal record on employment outcomes that appears substantially larger for African Americans. The sequence of interactions preceding hiring decisions suggests that black applicants are less often invited to interview, thereby providing fewer opportunities to establish rapport with the employer. Furthermore, employers' general reluctance to discuss the criminal record of an applicant appears especially harmful for black ex-offenders. Overall, these results point to the importance of rapport-building for finding work, something that the stigmatizing characteristics of minority and criminal status make more difficult to achieve.

Keywords: race; criminal record; discrimination; employment; low-wage labor markets

Roughly seven hundred thousand inmates are released from prison each year, a fivefold increase from the late 1970s (Bureau of Justice Statistics 2004). Consisting mostly of young men with less than a college education, about two-thirds of ex-prisoners remain out of work a year after prison release, and 60 percent are rearrested within three years (Bureau of Justice Statistics 2002; Petersilia 2003; Travis 2005). Those that can find steady work are less likely to return to prison and are better-equipped to assume the mainstream social roles of spouse and parent (Lopoo and Western 2005; Sampson and Laub 1993; Uggen 2000). Unfortunately, the goal of stable employment remains elusive for a large fraction of ex-offenders. The challenges of reentering society from prison are compounded for many by the racial stigma

NOTE: Research for this article was supported by grants from the National Science Foundation, the National Institute of Justice, and the JEHT Foundation. We gratefully acknowledge the generous assistance of Commissioner Patricia Gatling and the New York Commission on Human Rights and the contributions of Glenn Martin of the Fortune Society.

DOI: 10.1177/0002716208330793

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on May 5, 2009

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

produced by prejudice and discrimination. Black men are about six times more likely than whites to be sent to prison and are likewise overrepresented among released prisoners (Bureau of Justice Statistics 2004). Some evidence suggests that blacks may also pay a higher penalty for having a criminal record relative to otherwise similar whites (Pager 2007). Given these patterns, understanding the nature of criminal and racial stigma—and the combination of the two—represents an important goal for research and policy.

We study the effects of race and a prison record on employment with a large-scale field experiment conducted in New York City. In this study, teams of black and white men were matched and sent to apply for hundreds of low-wage jobs throughout the city, presenting equivalent resumés and differing only in their race and criminal background. These results build upon our earlier work (Pager 2003), pointing to a robust interaction between race and criminal background. Furthermore, this research allows us to look with more detail into the interpersonal contact between job seekers and employers for some insight into the process by which race and criminal background translate into labor market disadvantages.

We find a significant negative effect of a criminal record on employment outcomes, and one that appears substantially larger for African Americans. The sequence of interactions preceding hiring decisions suggests that black applicants are less often invited to interview, thereby providing fewer opportunities to establish rapport with the employer. Furthermore, employers' general reluctance to discuss the criminal record of an applicant appears especially harmful for black ex-offenders. Overall, these results point to the importance of rapport-building for finding work, something that the stigmatizing characteristics of minority and criminal status make more difficult to achieve.

Stigma, Rapport, and the Job-Matching Process

Little is known about the process by which employers select workers. Economic models of employment often assume that the productivity of prospective workers can be readily assessed, but in reality, employers often face acute information

---

Devah Pager is an associate professor of sociology and a faculty associate of the Office of Population Research at Princeton University. Her research focuses on institutions affecting racial stratification, including education, labor markets, and the criminal justice system. Her recent book, Marked: Race, Crime, and Finding Work in an Era of Mass Incarceration *(University of Chicago 2007), describes the results from a field experiment investigating the racial and economic consequences of large-scale imprisonment for contemporary U.S. labor markets.*

Bruce Western is a professor of sociology and director of the Program in Inequality and Social Policy at Harvard University. His research interests are in the areas of social stratification and inequality, political sociology, and statistical methods. His latest book, Punishment and Inequality in America *(Russell Sage Foundation 2006), examines the causes, scope, and consequences of the growth of the U.S. penal system.*

Naomi Sugie is a PhD student in sociology and social policy at Princeton University. Before beginning her graduate studies, she worked at the Vera Institute of Justice on child welfare, policing, and jail reentry issues. She is coauthor of a book profiling working mothers living in New York City shelters and is currently investigating the economic consequences of incarceration on families of offenders.

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on May 5, 2009

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

shortages in evaluating new hires. Particularly in low-wage job markets, where few concrete skills or experience are required, employers typically rely on limited information provided on a resumé or gathered during a cursory interview. Indeed, many employers claim to base hiring decisions on a "gut feeling" about candidates (Moss and Tilly 2001, 209), the source of which remains largely unknown.

Where employers are often looking for applicants with whom they feel an intuitive rapport, applicants with stigmatizing characteristics (such as minority status or a criminal background) may face special barriers to establishing such a rapport, even if possessing otherwise highly appealing characteristics. Indeed, a wealth of social psychological evidence indicates that negative stereotypes compromise interactions and undermine the ability of interaction partners to demonstrate traits that are inconsistent with stereotypical expectations. Experimental evidence suggests that people ask fewer questions of stereotyped targets (Trope and Thomson 1997) and selectively notice and retain information consistent with the stereotypes while ignoring information that is inconsistent with initial expectations (Fiske and Neuberg 1990). Although the effects of stereotypes have been shown to weaken as personalizing information becomes available (Allport 1954; Anderson 1999; Fiske and Neuberg 1990), perceivers are less likely to seek out or retain individuating information when confronted with members of stigmatized social groups.

It is easy to imagine how this process might play out in employment settings. In cases where employers are confronted with stereotyped applicants, they may be more likely to make negative attributions about the individual without probing deeper into the specific characteristics of the applicant in question. Employers may be less likely to grant an interview to such applicants and, during the course of an interview, may ask fewer questions or provide less opportunity for the applicant to present his or her profile in the best light (e.g., Word, Zanna, and Cooper 1974). While in some cases these dynamics may result from overt prejudice, they can also come from simple discomfort or more subtle, unconscious biases (Crocker, Major, and Steele 1998).

Most of the research on stereotyping and social interaction focuses on racial differences. In contemporary low-wage urban labor markets, a criminal conviction represents another source of disadvantage that may contribute to the differential treatment of young low-skill men. In fact, one might expect the effect of a criminal conviction to be more disqualifying for job applicants than racial stigma because of its direct association with negative behaviors—like dishonesty, violence, or unreliability—that suggest poor job performance on the job. On the other hand, criminal history is a legitimate topic of discussion in a job interview, with job applications commonly asking about criminal backgrounds and employers often discussing criminal convictions with job seekers. These opportunities for candid discussion may provide chances to defuse the effect of a criminal background, a strategy less available in the case of racial stigma, where prevailing norms discourage open conversations about race. Furthermore, relatively little is known about how various stigmatizing characteristics may interact to produce new forms of labor market disadvantage. How do employers' assumptions or concerns about black applicants overlap with or intensify their concerns about ex-offenders? In what ways do the barriers facing one applicant type (e.g., a black

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on May 5, 2009

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

applicant) contribute to the disadvantages experienced by those with additional stigmatizing characteristics (e.g., a black ex-offender)?

These ideas were previously examined in Pager's (2003) audit study of entry-level jobs in Milwaukee. Using an audit methodology, replicated in this article, two-person teams of trained testers were assigned resumés with equivalent schooling and work histories. Within each team, one tester was randomly assigned a criminal record. The applicant pairs applied for entry-level jobs, measuring the extent to which race and a criminal background represented barriers to employment. Milwaukee employers strongly disfavored job seekers with criminal records, and the penalty of the criminal record was especially large for blacks. These results suggested that minority status compounds the stigma of a criminal record, though the mechanism through which this stigma is exerted remained unobserved.

We replicate the design of the Milwaukee study in New York City, and look beyond the general patterns of employment to investigate the sequence of events that precede an ultimate hiring decision. In particular, we examine the patterns of interaction (quantity and kind) experienced by black and white job seekers and their relationship to ultimate hiring outcomes. By studying the hiring process through this lens, we can better understand how rapport-building is facilitated or compromised as a function of an applicant's race or criminal background.

Data and Methods

The New York City Hiring Discrimination Study sent matched teams of testers to apply for 250 real entry-level jobs throughout New York City over nine months in 2004. The testers were well-spoken, clean-shaven young men, ages twenty-two to twenty-six. Most were college-educated, between 5 feet 10 inches and 6 feet in height, and recruited in and around New York City. They were matched on the basis of their verbal skills, interactional styles (level of eye-contact, demeanor, and verbosity), and physical attractiveness. Testers were assigned fictitious resumés indicating identical educational attainment and comparability in quality of high school, work experience, and neighborhood of residence. Resumés were prepared in different fonts and formats and randomly varied across testers, with each resumé used by testers from each race group. Testers presented themselves as high school graduates with steady work experience in entry-level jobs. Finally, the testers passed through a common training program to ensure uniform behavior in job interviews. While in the field, they dressed similarly and communicated with teammates by cell phone to forewarn one another of unusual interview situations.

To study employers' responses to applicants with criminal records, we fielded two teams of testers. The first team paired two white applicants, one presenting a criminal record and the other a clean record. The second team paired two similar black applicants. None of the testers had real criminal backgrounds, but presented fictitious records to employers. Testers rotated which member of the pair presented criminal background information, which allowed for control of

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on May 5, 2009

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

within-pair differences that might affect hiring outcomes. The criminal record was typically disclosed in answer to the standard question on job applications: "Have you ever been convicted of a crime? If yes, please explain." When asked, testers were instructed to reveal that they were recently released from prison after serving eighteen months for a drug felony (possession with intent to distribute cocaine). Following Pager (2003), the tester's criminal record was also signaled on his resumé by listing work experience at a state prison and by listing a parole officer as a reference.[1]

For both teams, employers were sampled from job listings for entry-level positions, defined as jobs requiring no previous experience and no education greater than high school. Jobs included restaurant positions, retail sales, warehouse work, couriers, telemarketers, customer service positions, clerical workers, stockers, movers, delivery drivers, and a wide range of other low-wage positions. Job listings were randomly drawn each week from the classified sections of the *New York Times*, *Daily News*, *New York Post*, *Village Voice*, and craigslist. The broad range of job listings allows for extensive coverage of the entry-level labor market in New York. From the available population of job listings, we took a simple random sample of advertisements each week. Testers in each team applied to each job within a twenty-four-hour period, randomly varying the order of the applicants. Our dependent variable recorded positive responses in which a tester was either offered a job or called back for a second interview. Callbacks were recorded by voice mail boxes set up for each tester. For more information about the research design and methods, see Pager, Western, and Bonikowski (2009).

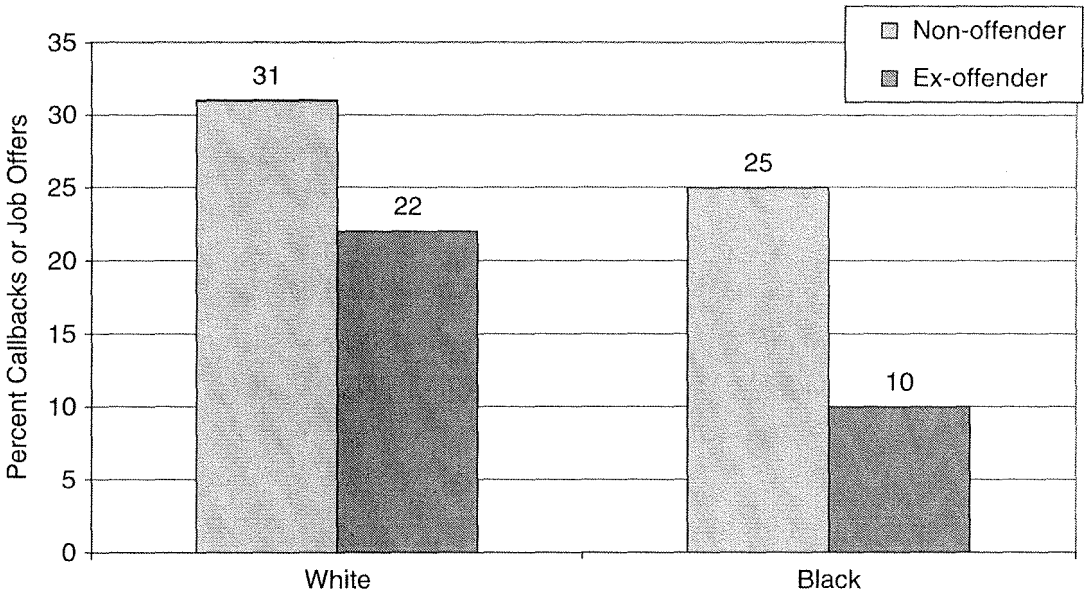
Results

Two key findings emerge from the audit results. First, as in earlier research, a criminal record has a significant negative impact on hiring outcomes, even for applicants with otherwise appealing characteristics. Across teams, a criminal record reduces the likelihood of a callback or job offer by nearly 50 percent (28 vs. 15 percent). Second, the negative effect of a criminal conviction is substantially larger for blacks than for whites. As shown in Figure 1, the magnitude of the criminal record penalty suffered by black applicants (60 percent) is roughly double the size of the penalty for whites with a record (30 percent). This interaction between race and criminal record is large and statistically significant, which indicates that the penalty of a criminal record is more disabling for black job seekers than whites. The intensification of the criminal record effect among blacks is consistent with earlier audit research (Pager 2007) and points to special barriers facing blacks in the transition from prison to work. Employers, already reluctant to hire blacks, appear particularly wary of blacks with known criminal histories. In the remainder of this article, we examine the sequence of interactions that lead to this ultimate pattern of results. As job applicants pass from the point of application to an interview, and from an interview to an offer, we witness some of the underlying dynamics that may shape employers' decision making and result in the systematic disadvantage of blacks with criminal records.

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on May 5, 2009

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

FIGURE 1
THE EFFECT OF RACE AND CRIMINAL BACKGROUND ON EMPLOYMENT OPPORTUNITIES



The importance of personal contact

One of the ways that job applicants can build rapport with employers is through the interview process. Though typically brief for low-wage jobs, interviews provide opportunities for applicants to demonstrate communication skills and commitment to work. For employers concerned about soft skills not reflected on a resumé, even a brief interaction can provide important information about the capacity of an applicant. Especially in the case of stigmatized applicants, personal contact may serve an important means of counteracting employers' initial stereotypes. As employers learn more about the person behind the category (e.g., ex-offender, black man), their comfort level with the applicant in question is likely to increase.

The evidence from our audit study indeed confirms that personal contact with an employer has a substantial impact on hiring outcomes. Restricting our sample to cases in which both team partners had the same level of contact with an employer, we find that testers who interact with employers are between four and six times more likely to receive a callback or job offer than those who do not; and personal contact reduces the effect of a criminal record by roughly 15 percent (see the appendix).[2] Personal contact thus seems to play an important role in mediating the effects of criminal stigma in the hiring process. At the same time,

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on May 5, 2009

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

this pathway to rapport-building may not be equally available to all applicant types. Although all testers in the study were instructed to request to speak to the person in charge of hiring and to proceed as far as they could in the interview process, some met with more success than others. In particular, race has a significant effect on the likelihood of personal contact between applicant and employer, with blacks roughly a third less likely to have the opportunity to interact with employers ($p < .001$). Employers appear to be screening on the basis of race in deciding whom to allow to proceed from application to interview (see Figure 2). By contrast, the effect of a criminal record has no discernable impact on the likelihood of interaction. Given that a criminal record is typically unobserved until an employer has spoken with the candidate and/or reviewed his materials, it is not surprising that this is typically not the basis of an employer's decision to interview.

Personal contact . . . seems to play an important role in mediating the effects of criminal stigma in the hiring process. At the same time, this pathway to rapport-building may not be equally available to all applicant types.

The barriers that emerge in this very early stage of hiring are likely consequential for the disparities observed. With fewer opportunities for face-to-face contact with employers, black applicants are limited in their ability to demonstrate specific skills and attributes. Particularly in the case of black ex-offenders, for whom employers' concerns are likely particularly strong, limits on interaction reduce opportunities to contextualize a conviction or to demonstrate evidence of successful rehabilitation.

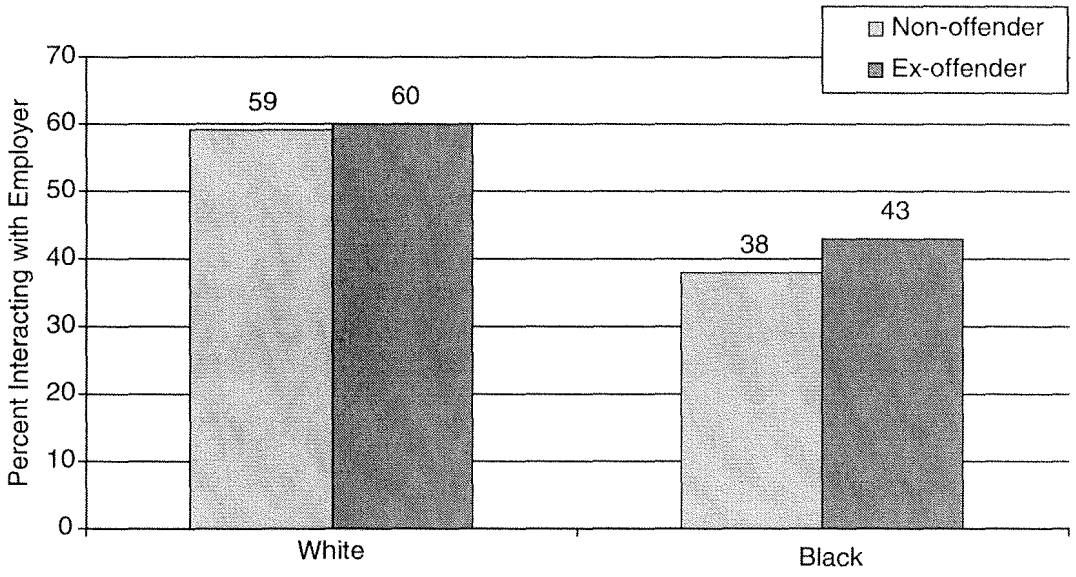
Quality of the interaction

Being granted an interview, or even a conversation, with the person in charge of hiring provides an important opportunity for establishing rapport. But the nature of that rapport also hinges upon what takes place during the interaction. Particularly for applicants with criminal records, the interview provides a key opportunity to assuage employers' concerns. In this respect, our testers had

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on May 5, 2009

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

FIGURE 2
THE LIKELIHOOD OF PERSONAL CONTACT WITH EMPLOYER, BY RACE AND CRIMINAL BACKGROUND



mixed experiences. The testers had a set script that they were instructed to convey to employers about their prior drug conviction and their commitment to rehabilitation. In some cases, employers were extremely receptive to discussing these issues, while in others, employers seemed uncomfortable or unwilling to broach the subject. In one interaction, for example, an employer inquired about the most recent job listed on Kevin's resumé, which was at a correctional facility. Kevin reports,

> I thought she was asking me what I did to get in there. I said, "It was for drug possession." She said, "No, not that. That's none of my business."

The employer then quickly moved on to discuss Kevin's previous work experience. It is unclear from this interaction what lingering doubts the employer may have had about Kevin's criminal background, but Kevin did not have the chance to explain further.

In another case, Anthony, an African American tester, reports,

> As she looked over [my resumé] I could barely hear her say, "Oh, I see." I don't know what it was in response to, but it was pretty quick so I would guess it was my conviction. . . . She then just looked up at me and said, "I'll give you a call." It seemed like she ended it a bit abruptly.

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on May 5, 2009

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Some employers seemed uncertain about what was legally or socially appropriate to ask about prior convictions, and others seemed simply uncomfortable with the topic or considered it outside of the realm of employment-related concerns. In these cases, it is difficult to interpret the employers' response to the criminal background, and the applicant typically had less opportunity to account for the stigmatizing record or address employers' underlying concerns.

In other cases, employers' concerns about the criminal record are more transparent. Worries about legal liability, for example, came up in this interview with Chad, an African American tester:

> When I finished the application I was interviewed by . . . a large white man with a thick mustache. He shook my hand, invited me to have a seat, and began to look over my resumé. He said, "First, I need you to explain this . . . correctional facility and parole officer reference." I told him that I was convicted of a drug charge—possession with intent to sell. "I can't hire you," he said. He went on to explain that a lot of things can happen and the liability is too great. He said, "Let's say you got into a fight with a guy and you were in the right. The police come and run your background, yes, now you're in the wrong, even though you may have been right. It wouldn't be good for you and it wouldn't be good for us. I couldn't hire you."

Sometimes, employers' negative reactions are less explicit, but their concern over the prospect of hiring an ex-offender is clear nevertheless. For example, Kevin, a white tester, reports his experience at a specialty foods store:

> I noticed a sign on the door which read, "Help wanted, part-time, some experience." . . . A few minutes later a man came out. . . . He told me that he had a great part-time position [and] there could be some full-time positions opening [in] a while. He pointed at my... application . . . and said, "Why did you write parole?" I said that I was currently on parole. [He] then looked me in the eye and said, "Did you commit a crime?" I said yes. He then looked down at the sheet and said that he really wasn't hiring right now.

Kevin's conviction record seemed to catch the employer by surprise. Within seconds, the many signals pointing to a job opportunity (help-wanted sign, "great part-time positions," etc.) disappeared, as the employer decided he was no longer hiring, or at least not hiring Kevin.

Despite these unpleasant experiences, not all employers reacted negatively upon noticing a criminal record. In fact, on a number of occasions, testers encountered extremely sympathetic employers. For example, Kevin records his experience in applying for a job at a car dealership:

> He saw the correctional facility [on my resumé] and said, "We're an Equal Opportunity Employer. We don't care about this. About 75 percent of the people in this business have a record anyway."

Kevin describes the end of the interview:

> He said he was going to call me. Then he said, "I know what you are thinking. This asshole is never going to call me. I will call you. Not because you have good sales experience but just because you need a chance."

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on May 5, 2009

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

This employer appears sympathetic to the plight of an ex-offender looking for work. In fact, the employer seems willing to privilege the desire to give Kevin a second chance over his need for workers with relevant experience. This employer called Kevin about the job two days later.

For employers who have ambivalent feelings about hiring ex-offenders, or who have anxieties about particular kinds of ex-offenders, interaction with the candidate allows the employer to interrogate these concerns directly.

To be sure, many of the sympathetic responses received by testers in the criminal record condition simply reflect the preexisting attitudes of employers, independent of the interaction. Employers who feel sympathetic toward ex-offenders are likely to express such sympathies in conversations with ex-offender applicants. But above and beyond employers' predispositions, we observe some evidence that the interaction itself can work to clarify and shape the employers' interpretation of the criminal record. For employers who have ambivalent feelings about hiring ex-offenders, or who have anxieties about particular kinds of ex-offenders, interaction with the candidate allows the employer to interrogate these concerns directly. In one case, for example, Keith, a white tester, has an extended conversation with the manager at a furniture rental store:

> After finishing the application I brought it back to [the employer], along with my resumé. He invited me to take off my backpack and have a seat. He began looking over the res./app. and his first question was, "Were you selling or using?" I told him, "Using. It was a minor thing. A stupid mistake and I'm now clean." He was sympathetic saying, "I gotcha. It was a question, not a criticism." [The employer then asked him a few questions about his driver's license and driving history.] He invited me to sit with him out at the door while he smoked. When we got there he informally sat me down and lit up. He turned to me and said, "So why should I hire you instead of one of the forty-seven other guys I got coming to me?" I told him, "I'll work hard for you, bust my ass. It's a condition of my parole that I work." He said, "But do you want to work?" I answered, "Yeah, I'm looking to get back into society . . ." He interrupted, "You want to get your shit together." "Yeah," I said. . . . The conversation ended with him saying, "My inclination is to say yes" (regarding hiring me). He added, "My bosses, the owners, are a little more close-minded than me. . . . Look, you paid, you don't have to pay for it the rest of your fucking life. People make mistakes. I'll get you my card."

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on May 5, 2009

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

This employer expresses some open-mindedness about Keith's criminal background from the start ("It was a question, not a criticism") but also wants evidence that taking a chance on Keith would be warranted amid the large pool of candidates. The conversation seems to provide important reassurances to the employer, who ends the interview with an encouraging note.

In another case, Anthony, an African American tester, discusses his criminal background with the manager of a health care company:

> [The employer] said, "I'm sure people must take double-takes [when they see the correctional facility]." I replied, "Yeah, that does happen sometimes." He told me that he knows the law, says I have to provide that information to a possible employer, but not to worry because he has had other employees who have "fucked up in the past." He said, "I feel safer knowing you're telling me up front than me having to wonder if you're gonna come here and tear shit up. Let's face it, interviews are bullshit. You can't know someone from a five-minute interview. So let's cut to the chase."
>
> "How long were you in?" (Eighteen months)
>
> "When'd you get out?" (A few weeks ago)
>
> "Ok, so you're fresh out and trying to get back on track?" (Yes, I am)
>
> "What'd you do?" (I had a small amount of drugs on me)
>
> "So you were guilty?" (Yes, I was young and made some mistakes but I learned from them and am completely drug-free)
>
> [Jokingly] "So you weren't innocent?" (No, it was my fault.)
>
> "Don't worry, I find that those that messed up and want to set things right are better workers."

In this interaction the employer does acknowledge certain concerns about hiring someone with a record but seems to respond favorably to Anthony's honesty and, after learning more of the details of Anthony's background, offers him an encouraging response.

Thus, while a criminal record has a significant negative impact on employment prospects of job seekers, some employers are willing to look beyond the conviction, or to downplay its significance in the context of other information acquired during the interview. In these cases, a kind of empathy seems to develop between employer and job seeker, with goodwill often translating into a substantial improvement in employment prospects. Of course, the types of individualizing information employers look to are not always in the applicants' control. Race and ethnicity, in particular, appear to affect some employers' interpretations of the seriousness of the criminal background and the depth of empathy generated by the interaction. For example, Keith, a white tester, reports on his interaction with the manager of a restaurant supply company:

> [The employer] sat me down and went over my application and resumé. He first saw [the correctional facility] and asked about my working there. I told him that I had been incarcerated. In a lower voice he said, "What did you do?" I told him about my being caught with cocaine, my time served, my current sobriety, which my parole officer could verify, the fact that it was a mistake and I had learned my lesson. At this point he said, "Zarriello . . . what is that, Italian?" I said, "Yes," and could immediately tell he was

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on May 5, 2009

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

> now on my side. He asked more questions like, "What happened exactly?" I told him I was in a car with some ex-friends that was pulled over and we were all searched and they found six grams in my possession. He asked, "You come from a nice Italian family? What do your parents think about this?" I told him they were disappointed but thought I had learned my lesson. He told me that "people make mistakes." He went on to say, "It would make me feel good to help a guy like you, more than just the rewards of doing my job, but good as a person if I can help someone."

This employer's emphasis on the value of second chances and the desire to help a young ex-offender get back on his feet are similar to sympathetic reactions we heard from other employers in interactions with both black and white testers. Here, though, we see the employer explicitly invoke Keith's ethnic background as the basis for solidarity and as a key turning point in the employer's reaction to the criminal background information, as Keith moves from being viewed as an ex-offender to someone from a "nice Italian family."

In a similar case, Kevin applied for a job with a staffing agency and was asked a number of detailed questions about his criminal history. Toward the end of the interview, Kevin reports,

> [The employer] said, "Do you have Irish in you?" I said, "On my mother's side I do." He asked what else I have. I said, "French." He was delighted! He said he has Irish and French in him, too. He said we could be related because we are both from FrIreland [France and Ireland]. I said yes and laughed with him.

Kevin concluded his report by offering his impression of the interaction: "He really wanted to help me out and seemed to be going to great lengths to find me a job."

Once again, the ethnic solidarity expressed in this interaction appears to help establish a positive rapport between candidate and employer. Conversations with employers thus simultaneously offer the opportunity to present personalizing information about the applicant's work ethic and commitment to rehabilitation but may also generate new bases for categorical distinctions.

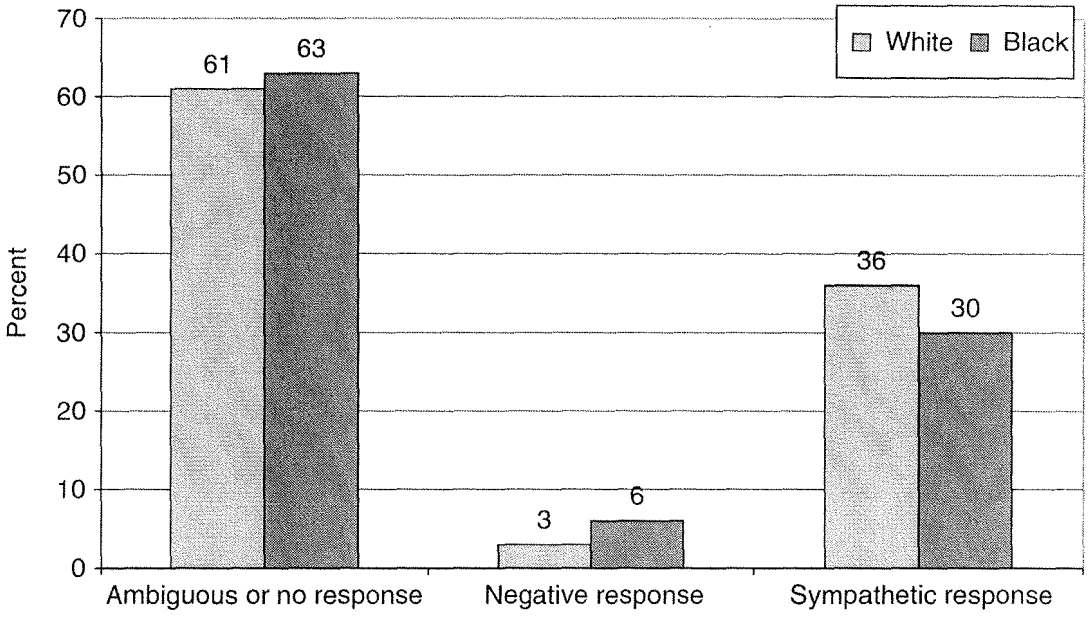
Employers thus appear to offer a range of reactions to ex-offender applicants, varying in terms of employers' comfort level in discussing criminal backgrounds and their evaluative assessments of this information. To examine employers' responses more systematically, we coded tester interactions with employers according to the nature of their response to the criminal record information, based on narrative data provided by the testers. Focusing on testers with criminal records who had personal contact with employers (roughly 50 percent of all tests), we code responses as "ambiguous or no response," "negative response," and "sympathetic response."

Looking to the results in Figure 3, we see that overall employers are most likely to avoid talking about the conviction altogether. Between two-thirds and three-quarters of employers either avoided the subject of the criminal record altogether or gave little indication of how they viewed the information. By contrast, less than 10 percent of employers made explicitly negative comments.

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on May 5, 2009

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

FIGURE 3
EMPLOYER REACTION TO CRIMINAL RECORD, BY RACE OF APPLICANT



Although the hiring outcomes from the audit study indicate a large negative impact of a criminal record on employers' evaluations, we see little of this reflected in their explicit comments to job applicants. If expressing a clear valence, employers are more likely to offer sympathetic reactions, with roughly 35 percent of employers coded as sympathetic toward the ex-offender applicant.[3] We see some evidence that blacks are more likely to receive a negative response (6 vs. 3 percent) and less likely to receive a sympathetic response (30 vs. 36 percent), though these differences are not statistically significant. Overall, these results point to a reluctance among employers to address the criminal record issue head-on, or to reveal their reaction to the record to the applicant in question. Our final question, then, considers the extent to which these differential responses in interaction correspond to differences in hiring outcomes.

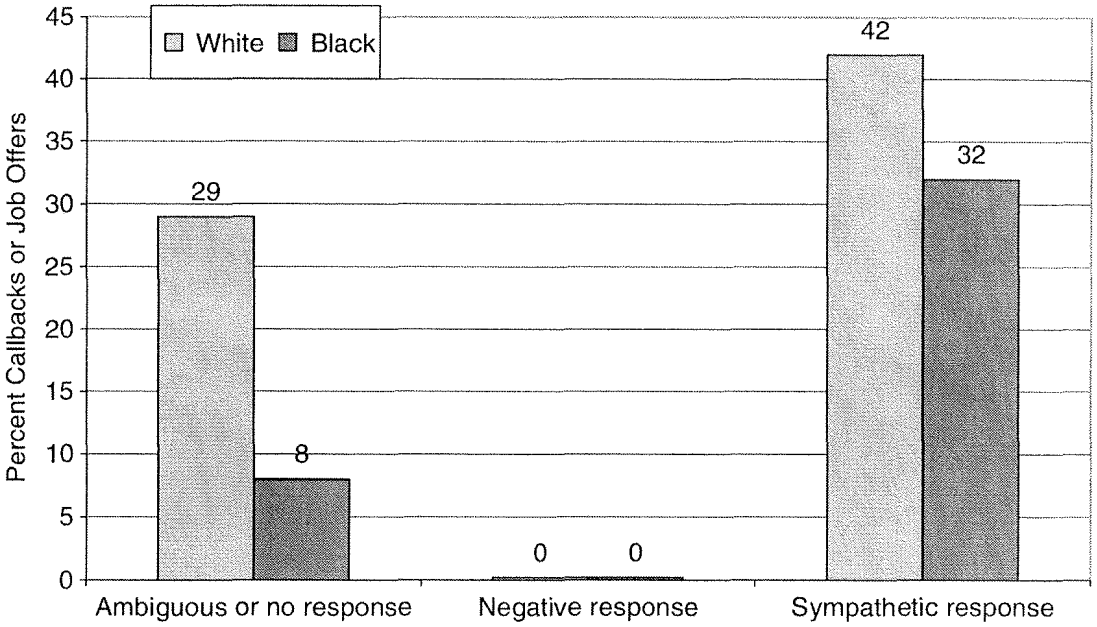
Relationship between type of interaction and employment outcomes

The nature of interaction between employer and applicant is significant primarily to the extent that it proves consequential for hiring. Matching interaction experiences with employment outcomes provides some leverage on the pathways through which ex-offenders find opportunity. Figure 4 presents the percentage

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on May 5, 2009

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

FIGURE 4
THE LIKELIHOOD OF A CALLBACK OR JOB OFFER, BY RACE AND INTERACTION TYPE



of applicants with criminal records who received a callback or job offer, by race of the applicant and type of employer response. Not surprisingly, employment outcomes are most favorable among those who received a sympathetic response from employers. These employers are not simply paying lip service to the value of second chances but demonstrate an actual willingness to hire ex-offenders. Among those who receive sympathetic responses from employers, whites are more likely to receive an actual callback or job offer (42 vs. 32 percent), although this difference is not statistically significant.

Showing even more consistency between interactions and outcomes, employers who express negative reactions to applicants with criminal records in no cases made offers or callbacks to these applicants. The group with less consistent results includes employers who offer no reaction or ambiguous reactions to the criminal background. Among these employers, we see a large racial difference in outcomes, with white applicants roughly three times more likely to receive a callback or job offer relative to blacks who have similar encounters (29 vs. 8 percent). Relative to those who receive a sympathetic response, the penalty associated with limited or no discussion about the criminal record is roughly 30 percent

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on May 5, 2009

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

for whites; for blacks, this limited interaction appears far more consequential, resulting in 75 percent fewer callbacks or job offers relative to those who received a sympathetic reaction. Though we cannot directly interpret employers' underlying reactions, this evidence is consistent with the role of stereotypes inhibiting the acquisition and impact of personalizing information. If employers who are concerned about the record among black applicants choose to remain silent about the issue, the applicant then has little opportunity to anticipate or address the employer's concerns. Where for white ex-offenders this reduced communication does not appear overly consequential, black ex-offenders seem to face substantially lower employment prospects as a result.

Conclusion

The results of this study show a strong reluctance among employers to hire applicants with criminal records, especially when considering black ex-offenders. Despite the many appealing personal characteristics of our testers, employers often appear to base their decisions on the more salient markers of race and criminal background. What is perhaps more noteworthy in these data relates to the cases in which testers with criminal backgrounds *are* given a chance at employment. Employment prospects improve significantly for applicants who have a chance to interact with the hiring manager, and more so among those who elicit sympathetic responses in the course of those interactions. Surely, some of this variation is attributable to preexisting characteristics and preferences of the employers, with little or no effect of personal contact. Employers who are eager to hire will be more likely to meet with applicants on the spot, and those who are sympathetic toward ex-offenders will be more likely to express such sentiments in the course of interaction. Still, we suspect that the interaction itself plays a nontrivial role in this hiring process. Employers have many reasons to be concerned about applicants with prior histories of incarceration. Concerns about theft, violence, and drug use are all relevant, not to mention the more mundane concerns over worker reliability and performance. Personal contact with an applicant cannot reveal all of these issues but can help to provide some signals as to the disposition of the applicant and can help the employer develop a "gut feeling" about whether this individual is likely to diverge from the stereotype of the ex-con.

Unfortunately, the ability to have such a hearing does not appear available to all applicants. Blacks are significantly less likely to be invited to interview by employers, offering them fewer opportunities to present indicators at odds with their stigmatized group membership. Furthermore, although the distribution of reactions from employers is roughly similar among black and white applicants with criminal backgrounds, actual employment outcomes differ for those who have little opportunity to discuss their criminal record: among whites, these limited interactions are not overly consequential; whereas for blacks, job opportunities appear substantially reduced.

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on May 5, 2009

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

These findings must be contextualized in light of the sampling design of the study, which focused exclusively on jobs obtained through formal classified listings. Given that many job seekers find employment through social networks and other informal channels, our analysis may understate opportunities for personal contact made possible through mediated contacts. However, evidence on social networks in employment suggests racialized consequences of these pathways as well, with blacks less likely to obtain quality leads to employment from their networks relative to similarly situated whites (cf. Royster 2003). Racial disparities in access to social networks have also been shown in the case of ex-offenders (Sullivan 1989). These informal methods of job search behavior, therefore, may result in greater evidence of racial disparities in employment following incarceration than what is reported here.

Overall, these findings point to the importance of rapport-building in the employment process, particularly for applicants with stigmatizing characteristics. In light of these findings, policy intervention should aim to defuse stigma and provide employers with more information about their prospective workers. Initiatives that facilitate the matching of workers with employers in ways that help to overcome these initial barriers may have a substantial impact. Job referral services that act as labor market intermediaries who vouch for job applicants represent one important policy approach to bridging this divide. Certificates of rehabilitation and public education campaigns might also weaken the effects of stereotyping.

As incarceration rates have increased over the past few decades, official criminality compounds the stigma of race and deepens the economic disadvantage of young African American men. Instead of merely adding to the deficits of low-skill black men, a criminal record modifies the effect of racial discrimination, which raises the bar to employment higher for blacks than similarly situated whites. In this context, we can understand the growth of incarceration rates, and the racial disparities that characterize them, as producing a new form of institutional racism with wide-reaching economic effects.

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on May 5, 2009

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Downloaded from http://anm.sagepub.com at PRINCETON UNIV LIBRARY on May 5, 2009

APPENDIX

THE EFFECT OF PERSONAL CONTACT ON THE LIKELIHOOD OF A CALLBACK/JOB OFFER AND THE MAGNITUDE OF CRIMINAL STIGMA

	No Interaction with Employer				Both Testers Interacted with Employer				
	No Record (NR)	Criminal Record (CR)	Ratio 1 (NR/CR)	*n*	NR	CR	Ratio 2 (NR/CR)	*n*	Percentage Change in CR Effect
White	8.82	5.88	1.50	34	44.44	35.19	1.26	54	–0.16
Black	10.00	2.82	3.55	70	41.86	13.95	3.00	43	–0.15

NOTE: First two columns of each section represent rates of positive response (callback/job offer) for testers by race and criminal status. Percentage change in the effect of a criminal record (final column) is calculated as: (ratio 1 – ratio 2)/ratio 1.

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Notes

1. Results from Pager (2003) suggest that providing information about a criminal record to employers who do not request the information does little to affect hiring decisions. Those employers who request the information are those most likely to use it.

2. We restrict our sample here to cases in which both testers on a team received the same level of personal contact to better control for compositional differences between those employers more or less likely to interview candidates on the spot. By comparing the effect of a criminal record within teams where either both or neither tester interacts with the employer, we can better understand the ways in which personal contact may mediate the effects of stigma. This sample restriction has little effect on the substantive conclusions of the analysis.

3. This proportion corresponds closely with the 33 percent of urban employers surveyed by Holzer (1996, 59) who report that they would "probably accept" or "definitely accept" an applicant with a criminal background.

References

Allport, Gordon. 1954. *The nature of prejudice*. New York: Doubleday Anchor.

Anderson, Elijah. 1999. *Code of the streets: Decency, violence, and the moral life of the inner city*. New York: Norton.

Bureau of Justice Statistics. 2002. *Recidivism of prisoners released in 1994*. By Patrick Langan and David Levin. Washington, DC: U.S. Department of Justice.

———. 2004. *Reentry trends in the U.S.* Washington, DC: U.S. Department of Justice. http://www.ojp.usdoj.gov/bjs/reentry/reentry_contents.htm#contents.

Crocker, Jennifer, Brenda Major, and Claude Steele. 1998. Social stigma. In *The handbook of social psychology*, ed. D. Gilbert, S. Fiske, and G. Lindzey, 504-53. Boston: McGraw-Hill.

Fiske, Susan, and S. L. Neuberg. 1990. A continuum model of impression formation: from category-based to individuating processes: Influence of information and motivation on attention and interpretation. In *Advances in experimental psychology*, ed. M. P. Zanna, 1-74. New York: Academic Press.

Holzer, Harry. 1996. *What employers want: Job prospects for less educated workers*. New York: Russell Sage Foundation.

Lopoo, Leonard M., and Bruce Western. 2005. Incarceration and the formation and stability of marital unions. *Journal of Marriage and Family* 67:721-34.

Moss, Philip, and Chris Tilly. 2001. *Stories employers tell: Race, skill, and hiring in America*. New York: Russell Sage Foundation.

Pager, Devah. 2003. The mark of a criminal record. *American Journal of Sociology* 108:937-75.

———. 2007. *Marked: Race, crime, and finding work in an era of mass incarceration*. Chicago: University of Chicago Press.

Pager, Devah, Bruce Western, and Bart Bonikowski. 2008. *Race at work: Results from a field experiment of discrimination in low wage labor markets*. Princeton, NJ: Princeton University Press.

Petersilia, Joan. 2003. *When prisoners come home: Parole and prisoner reentry*. New York: Oxford University Press.

Royster, Deirdre. 2003. *Race and the invisible hand: How white networks exclude black men from blue-collar jobs*. Berkeley: University of California Press.

Sampson, Robert J., and John H. Laub. 1993. *Crime in the making: Pathways and turning points through life*. Cambridge, MA: Harvard University Press.

Sullivan, Mercer L. 1989. *"Getting paid": Youth crime and work in the inner city*. Ithaca, NY: Cornell University Press.

Travis, Jeremy. 2005. *But they all come black: Facing the challenges of prisoner reentry*. Washington, DC: Urban Institute Press.

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on May 5, 2009

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Trope, Yaacov, and Erik P. Thomson. 1997. Looking for truth in all the wrong places? Asymmetric search of individuating information about stereotyped group members. *Journal of Personality and Social Psychology* 73:229-41.

Uggen, Christopher. 2000. Work as a turning point in the life course of criminals: A duration model of age, employment, and recidivism. *American Sociological Review* 65:529-46.

Word, Carl O., Mark P. Zanna, and Joel Cooper. 1974. The nonverbal mediation of self-fulfilling prophecies in interracial interactions. *Journal of Experimental Social Psychology* 10:109-20.

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on May 5, 2009

Discrimination in a Low-Wage Labor Market: A Field Experiment

Devah Pager
Princeton University

Bruce Western
Harvard University

Bart Bonikowski
Princeton University

Decades of racial progress have led some researchers and policymakers to doubt that discrimination remains an important cause of economic inequality. To study contemporary discrimination, we conducted a field experiment in the low-wage labor market of New York City, recruiting white, black, and Latino job applicants who were matched on demographic characteristics and interpersonal skills. These applicants were given equivalent résumés and sent to apply in tandem for hundreds of entry-level jobs. Our results show that black applicants were half as likely as equally qualified whites to receive a callback or job offer. In fact, black and Latino applicants with clean backgrounds fared no better than white applicants just released from prison. Additional qualitative evidence from our applicants' experiences further illustrates the multiple points at which employment trajectories can be deflected by various forms of racial bias. These results point to the subtle yet systematic forms of discrimination that continue to shape employment opportunities for low-wage workers.

Despite legal bans on discrimination and the liberalization of racial attitudes since the 1960s, racial differences in employment remain among the most enduring forms of economic inequality. Even in the tight labor market of the late 1990s, unemployment rates for black men remained twice that for whites. Racial inequality in total joblessness—including those who exited the labor market—increased among young men during this period (Holzer and Offner 2001). Against this backdrop of persistent racial inequality, the question of employment discrimination has generated renewed interest. Although there is much research on racial disparities in employment, the contemporary relevance of discrimination remains widely contested.

One line of research points to the persistence of prejudice and discrimination as a critical factor shaping contemporary racial disparities (Darity and Mason 1998; Roscigno et al. 2007). A series of studies relying on surveys and in-depth interviews finds that firms are reluctant to hire young minority men—especially blacks—because they are seen as unreliable, dishonest, or lacking in social or cognitive skills (Holzer 1996; Kirschenman and Neckerman 1991; Moss and Tilly 2001; Waldinger and Lichter 2003; Wilson 1996: chap. 5). The strong negative attitudes expressed by employers suggest that race remains highly

Direct all correspondence to Devah Pager, Department of Sociology, Princeton University, Princeton, NJ 08544 (pager@princeton.edu). This research has been supported by grants from the National Science Foundation, the National Institute of Justice, the JEHT Foundation, the Princeton Research Institute on the Region, and the Industrial Relations Section of Princeton University. The first author acknowledges generous support from NSF CAREER, NIH K01, and a W.T. Grant Scholars Award. We also gratefully acknowledge the support of the New York City Commission on Human Rights, and Commissioner Patricia Gatling. Thanks to Glenn Martin, Don Green, and the many workshop participants who provided generous feedback on earlier drafts of this article.

American Sociological Review, 2009, Vol. 74 (October:777–799)

From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

salient in employers' evaluations of workers. At the same time, research relying on interviews with employers leaves uncertain the degree to which self-reported attitudes are influential in actual hiring decisions (Pager and Quillian 2005). Indeed, Moss and Tilly (2001:151) report the puzzling finding that "businesses where a plurality of managers complained about black motivation are more likely to hire black men." In fact, across a series of analyses controlling for firm size, starting wage, the percent black in the relevant portion of the metropolitan area, and a business's average distance from black residents in the area, Moss and Tilly find that employers who overtly criticize the hard skills or interaction skills of black workers are between two and four times more likely to hire a black worker (pp.151–52). Hiring decisions, of course, are influenced by a complex range of factors, racial attitudes being only one. Employers' stated preferences do not provide a clear picture of the degree to which negative attitudes about blacks translate into active forms of discrimination.

Research focusing on wages rather than employment offers even less evidence of contemporary discrimination. Neal and Johnson (1996), for example, estimate wage differences between white, black, and Latino young men. They find that two thirds of the black-white gap in wages in 1990 to 1991 can be explained by race differences in cognitive test scores measured 11 years earlier, and test scores fully explain wage differences between whites and Latinos. This and similar studies trace the employment problems of young minority men primarily to skill or other individual deficiencies, rather than any direct effect of discrimination (Farkas and Vicknair 1996; Neal and Johnson 1996; O'Neill 1990). Heckman (1998:101–102) puts the point most clearly, writing that "most of the disparity in earnings between blacks and whites in the labor market of the 1990s is due to differences in skills they bring to the market, and not to discrimination within the labor market." He goes on to describe labor market discrimination as "the problem of an earlier era."

Does employer discrimination continue to affect labor market outcomes for minority workers? Clear answers are elusive because discrimination is hard to measure. Without observing actual hiring decisions, it is difficult to assess exactly how and under what conditions race shapes employer behavior. We address this issue with a field experiment that allows direct observation of employer decision making. By presenting equally qualified applicants who differ only by race or ethnicity, we can observe the degree to which racial considerations affect real hiring decisions. Furthermore, we move beyond experimental estimates of discrimination to explore the *processes* by which discrimination occurs. Examining the interactions between job seekers and employers, we gain new insights into how race influences employers' perceptions of job candidate quality and desirability. Studying the multifaceted character of discrimination highlights the range of decisions that collectively reduce opportunities for minority candidates.

CONCEPTUALIZING DISCRIMINATION

Empirical studies often portray discrimination as a single decision. Research on employment disparities, for example, considers the role of discrimination at the point of initial hire; research on pay disparities considers discrimination at the point of wage-setting decisions. In reality, discrimination may occur at multiple decision points across the employment relationship. In this way, even relatively small episodes of discrimination—when experienced at multiple intervals or across multiple contexts—can have substantial effects on aggregate outcomes.

Depictions of discriminators also often portray the labor market as divided neatly between employers with a "taste for discrimination" and those who are indifferent to race (Becker 1957). Consequently, it is suggested, job seekers can avoid discrimination by sorting themselves into sectors of the labor market where discrimination is less likely to occur (Heckman 1998:103). Fryer and Levitt (2003:5) characterize employers according to a similar dichotomy, with applicants best advised to identify and avoid employers prone to discrimination, rather than wasting time pursuing job opportunities among firms unwilling to hire them: "In the face of discriminatory employers, it is actually in the interest of both employee and employers for Blacks to signal race, either via a name or other résumé information, rather than undertaking a costly

From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

interview with little hope of receiving a job offer." According to this conceptualization of labor market discrimination, racial preferences or biases are fixed and concentrated among a specific subset of employers.

Other evidence challenges this tidy distinction between employers who do and do not discriminate. Alternative formulations of labor market discrimination encourage us to view the process as more interactive, contextual, and widespread. Theories of both statistical discrimination and stereotypes view race as a heuristic employers use to evaluate job applicants about whom little is known. Here, group-based generalizations provide guidance about the expected profile of individuals from a given group and facilitate decision making when information or time are scarce (Aigner and Cain 1977; Fiske 1998). Heuristics of this kind are pervasive (and often unconscious). Their effects may vary depending on the availability of and attention to person-specific information (such as that conveyed through application materials or in an interview) that may interact with and potentially override initial expectations.

A long line of social psychological research investigates how stereotypes give way to individualizing information, as well as the conditions under which stereotypes demonstrate a stubborn resistance to change (Bodenhausen 1988; Fiske 1998; Trope and Thomson 1997).[1] This research suggests that salient personalizing information can quickly counteract stereotyped expectations; however, in evaluating difficult-to-observe or ambiguously relevant characteristics, or when decision makers have competing demands on their attention, stereotypes often filter information in ways that preserve expectations (Darley and Gross 1983; Dovidio and Gaertner 2000; Gilbert and Hixon 1991). In these cases of decision making under uncertainty, racial preferences or biases are unlikely to be expressed in any static or uniform way, but will vary in intensity and consequence depending on other characteristics of the applicant, the employer, and the interaction between the two.

[1] Theories of statistical discrimination also predict employer responsiveness to individual characteristics (e.g., Altonji and Pierret 2001; Oettinger 1996; cf. Pager and Karafin 2009).

In addition to noting the varying role of race across employment interactions, some research shifts the focus from employer characteristics to the characteristics of the job for which a given worker is being considered. Previous research points to the negative consequences of the changing composition of low-wage jobs for black men, with the shift from manufacturing to services skewing the distribution of skill demands toward "soft skills," for which black men are considered lacking (Moss and Tilly 2001). Jobs involving customer service or contact with clients heighten the salience of race because of employers' concerns about the dress and demeanor of young black men (Moss and Tilly 2001). Jobs at the "back of the house" or those emphasizing manual skills are less likely to activate concerns of this kind. In this scenario, discrimination may obtain not at the employer level but at the job level, with black applicants excluded from some job types and channeled into others. In this case, we would look to variation in discrimination not among employers but among the job openings for which workers are being considered.

Rather than viewing discrimination as a single decision, or as the result of a small group of highly prejudiced employers, a growing body of research points to the variable contexts that shape how information about applicants may be filtered and interpreted along racial lines. Decision making under uncertainty and the race-typing of jobs both make discrimination more likely. To capture the contingent and cumulative effects of discrimination implied by these theories requires an examination of how experiences of discrimination may be distributed across a wide range of decision points and may vary depending on interactions among the employer, the applicant, and the job in question.

THE CHANGING LANDSCAPE OF LOW-WAGE LABOR MARKETS

Economic theory predicts the decline of discrimination through market competition (Becker 1957), but several features of contemporary low-wage labor markets may sustain or renew racialized decision making. Shifts in the composition of both low-wage jobs and workers have potentially created new incentives and opportunities for employers to enact racial preferences in hiring. First, low-wage job growth is

From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

concentrated in service industries, in positions that place a heavy emphasis on self-presentation, interaction with customers, and other personality-related attributes (Moss and Tilly 2001). As discussed earlier, employers consistently express concerns over the "soft skills" of black men, implying a potential skills mismatch between the skill requirements of new job growth and the perceived skill profile of black male job seekers. Furthermore, because many of the qualities valued by employers for contemporary low-wage jobs are difficult to evaluate from a written application or brief meeting, generalized negative perceptions of minority workers may be more difficult for individual minority applicants to disconfirm (Biernat and Kobrynowicz 1997).

Second, low-wage labor markets today are characterized by increasing heterogeneity of the urban minority work force, with low-skill black workers now more likely to compete with other minority groups—in particular, low-skill Latino workers. Interviews with employers in Los Angeles and Chicago suggest consistent preferences for Latinos over blacks, with Latino workers viewed as more pliant, reliable, and hard-working (Kirschenman and Neckerman 1991; Waldinger and Lichter 2003). Given these racial preferences among employers, growing competition within the low-wage labor market may leave black men vulnerable to discrimination relative not only to whites, but to Latinos as well.

Finally, low-wage labor markets are increasingly supplied by workers with criminal records. Nearly a third of black men without a college degree have prison records by their mid-30s, adding to employers' reservations about black male job applicants (Pager 2007b; Pettit and Western 2004). The high rate of incarceration makes a criminal record a newly important source of stigma that is worth studying in its own right. Moreover, we can view a criminal record as an extreme and authoritative signal of the kinds of problematic behaviors that employers ascribe to young black men. In this context, separating the effects of criminal stigma from race provides a useful benchmark for measuring racial stigma. In the first effort in this direction, Pager's (2003) research in a Milwaukee field experiment compared racial and criminal stigma among matched pairs of job seekers. Fielding a pair of black and a pair of white job applicants (in which one member of each pair was randomly assigned a criminal record), Pager found that a black applicant with no criminal background experiences job prospects similar to those of a white felon. That blackness confers the same disadvantage as a felony conviction helps calibrate the deeply skeptical view of young black men in the eyes of Milwaukee employers.

The growing importance of soft skills, ethnic heterogeneity, and job seekers with criminal records suggest the persistence or increasing incidence of discrimination in contemporary low-wage labor markets. Whether based on statistical generalizations or inaccurate stereotypes, preconceived notions about the characteristics or desirability of black men relative to other applicant types are likely to structure the distribution of opportunity along racial lines.

METHODS FOR STUDYING LABOR MARKET DISCRIMINATION

Racial discrimination in the labor market is typically studied by comparing the wages of whites and minorities, statistically controlling for human capital characteristics. Estimates from a variety of social surveys suggest that the black-white difference in hourly wages among men usually range between about 10 and 20 percent (Cancio, Evans, and Maume 1996; Darity and Meyers 1998; Neal and Johnson 1996). Although widely used, this residual method, in which discrimination is defined as the unexplained race difference in wages, is sensitive to the measurement of human capital. Where race differences in human capital are incompletely observed, the effect of discrimination may be overestimated (Farkas and Vicknair 1996; Neal and Johnson 1996).

Residual estimates of discrimination infer employer behavior from data on workers' wages. Field experiments, by contrast, offer a more direct approach to the measurement of discrimination. This approach, also referred to as an audit methodology, involves the use of matched teams of job applicants—called testers—who apply to real job openings and record responses from employers. Testers are assigned equivalent résumés and are matched on a variety of characteristics like age, education, physical appearance, and interpersonal skills.

From callbacks to labor markets - (in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Because black and white testers are sent to the same firms, and testers are matched on a wide variety of characteristics, much of the unexplained variation that confounds residual estimates of discrimination is experimentally controlled.

In part due to taxing logistical requirements, the use of in-person audit studies of employment remains rare, with only a handful of such studies conducted over the past 20 years (Bendick, Jackson, and Reinoso 1994; Bendick et al. 1991; Cross et al. 1990; Pager 2003; Turner, Fix, and Struyk 1991).[2] Moreover, the typical emphasis on a single comparison group leaves several significant features of contemporary urban labor markets unexplored.

By studying both race and criminal background, the Milwaukee audit study represents an important starting point for this project (Pager 2003). The Milwaukee study examined the influence of the criminal justice system on labor market stratification by studying the effect of a criminal record for black and white job seekers. Although race emerged as a key theme in the study's findings, the topic of racial discrimination was not a central focus. Moreover, the research design yielded only indirect evidence of racial discrimination because black and white testers did not apply to the same employers. Our ability to investigate when and how racial discrimination occurs is therefore limited in this context.

The current study updates and extends earlier research in several ways. First, we focus directly on the question of racial discrimination, in both conceptualization and design. This emphasis allows us to situate our research within ongoing debates about discrimination and to provide a rigorous design for detecting racial discrimination. Second, we move beyond standard two-race models of discrimination by including matched black, white, and Latino job seekers, reflecting the racial heterogeneity of large urban labor markets. To our knowledge, this is the first study of its kind to simultaneously examine the employment experiences of three racial/ethnic groups. Third, to help calibrate the magnitude of racial preferences, we compare applicants affected by varying forms of stigma; specifically, we compare minority applicants with white applicants just released from prison. Where the Milwaukee study attempted this comparison across teams, the present analysis provides a direct test by comparing the outcomes of minority and ex-offender applicants who visit the same employers. Finally, we extend our analysis from the quantitative evidence of differential treatment to a rich set of qualitative data that allow for an exploration of the process of discrimination. Drawing from the testers' extensive field notes that describe their interactions with employers, we provide a unique window into the range of employer responses that characterize discrimination in contemporary low-wage labor markets.

RESEARCH DESIGN AND METHODS

The New York City Hiring Discrimination Study sent matched teams of testers to apply for 340 real entry-level jobs throughout New York City over nine months in 2004. The testers were well-spoken, clean-cut young men, ages 22 to 26. Most were college-educated, between 5 feet 10 inches and 6 feet in height, and recruited in and around New York City. They were matched on the basis of their verbal skills, interactional styles (level of eye contact, demeanor, and verbosity), and physical attractiveness. Testers were assigned fictitious résumés indicating identical educational attainment and comparable qualities of high school, work experience (quantity and kind), and neighborhood of residence. Résumés were prepared in different fonts and formats and randomly varied across testers, with each résumé used by testers from each race group. Testers presented themselves as high school graduates with steady work experience in entry-level jobs. Finally, the testers passed a common training program to ensure uniform behavior in job interviews. While in the field, the testers dressed similarly and communicated with teammates by cell phone to anticipate unusual interview situations.

We fielded two teams that each included a white, Latino, and black tester. To help ensure

[2] For a summary of the results of earlier audit studies of employment, see Heckman and Siegelman (1993) and Pager (2007a). Correspondence studies, which rely on résumés sent by mail rather than in-person applications (e.g., Bertrand and Mullainathan 2004), are less costly but rely on application procedures less suited for low-wage labor markets (which typically require in-person applications).

From callbacks to labor markets - In-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

comparability, the Latino testers spoke in unaccented English, were U.S. citizens of Puerto Rican descent, and, like the other testers, claimed no Spanish language ability. The first team tests a standard racial hierarchy, with the white tester serving as a benchmark against which to measure variation in racial and ethnic discrimination. To calibrate the magnitude of racial stigma, the second team compares black and Latino testers with a white tester with a criminal record. The criminal record was typically disclosed in answer to the standard question on employment applications, "Have you ever been convicted of a crime? If yes, please explain." We instructed testers to reveal, when asked, that they had recently been released from prison after serving 18 months for a drug felony (possession with intent to distribute, cocaine). In addition, following Pager (2003), the white tester's criminal record was also signaled on the résumé by listing work experience at a state prison and by listing a parole officer as a reference.[3]

For both teams, we sampled employers from job listings for entry-level positions, defined as jobs requiring little previous experience and no more than a high school degree. Job titles included restaurant jobs, retail sales, warehouse workers, couriers, telemarketers, customer service positions, clerical workers, stockers, movers, delivery drivers, and a wide range of other low-wage positions. Each week, we randomly drew job listings from the classified sections of the *New York Times, Daily News, New York Post, Village Voice,* and the online service Craigslist. The broad range of job listings allowed for extensive coverage of the entry-level labor market in New York. From the available population of job listings, we took a simple random sample of advertisements each week. Testers in each team applied to each job within a 24-hour period, randomly varying the order of the applicants.

Our dependent variable records any positive response in which a tester was either offered a job or called back for a second interview. We recorded callbacks using voicemail boxes set up for each tester. For employer i ($i = 1, \ldots, N$) and tester t ($t = W, B,$ or L for white, black, or Latino), a positive response, y_{it}, is a binary variable that scores 1 for a job offer or callback, and 0 otherwise. We define the level of differential treatment as the ratio in positive response rates for each comparison, $r_{WB} = \bar{y}_W / \bar{y}_B$, where $\bar{y}_t$ is the proportion of positive responses for testers of race t. Under the null hypothesis of equal treatment, $r_{WB} = 1$, the proportion of positive responses received by each racial group is equal. For data on matched pairs, several statistical tests have been proposed that use within-pair comparisons to account for the correlation of observations from the same pair (e.g., Agresti 1990; Heckman and Siegelman 1993). In our case, where three testers are sent to the same employer, we have a matched triplet and information from all three testers should ideally contribute to an inference about a contrast between any two. Ghosh and colleagues (2000) suggest that matched pairs can be fit with a hierarchical logistic regression with a random effect for each pair. We generalize their approach to our matched triplets, fitting a random effect for each employer. If the probability of a positive response is given by $E(y_{it}) = p_{it}$, the hierarchical model is written

$$\log\left(\frac{p_{it}}{1-p_{it}}\right) = \alpha_i + \beta B_{it} + \gamma L_{it},$$

where B_{it} is a dummy variable for blacks, L_{it} is a dummy variable for Latinos, and the random effects for employers, α_i, is given a normal distribution. The employer effects, α_i, induce a correlation among observations from the same employers and reduce standard errors, as in the usual matched-pair inference. We estimate the models with Markov Chain Monte Carlo methods. We construct intervals for the ratios (r_{WB}, r_{WL}, and r_{BL}) by taking random draws from the posterior predictive distribution of y_{it}. Alternative methods that adjust for clustering by employer yield similar results to those reported below.

The Problems of Matching

The quality of audit results depends on the comparability of the testers. Because race cannot be experimentally assigned, researchers must rely on effective selection and matching to construct audit teams in which all relevant characteristics

[3] Results from Pager (2003) suggest that providing information about a criminal record to employers who do not request the information does little to affect hiring decisions.

From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

of testers are similar—something that may leave substantial room for bias. Heckman and Siegelman (1993) argue that researchers know little about the hard-to-observe characteristics highly prized by employers. If testers are poorly matched, evidence of discrimination may be merely an artifact of idiosyncratic tester characteristics.

Bertrand and Mullainathan (2004) remove tester effects in a "correspondence test" that sent résumés with common white and black names to employers in Boston and Chicago. Their design allows the random assignment of résumé characteristics to white- and black-sounding names, largely removing concerns about unobserved characteristics. Résumés with white names were 50 percent more likely than those with black names to receive callbacks from employers (9.7 versus 6.5 percent). Studies of this kind provide some reassurance that results from the body of audit research are not driven by tester effects alone.

Because we rely on in-person audits for our study of low-wage labor markets, the effective matching of testers is a key concern.[4] We reviewed more than 300 applicants to identify our final team of 10 testers.[5] Successful applicants were subject to two lengthy screening interviews and a written test, a far more probing job selection process than the testers encountered in their fieldwork.[6] Each tester passed a standard training period, was required to dress uniformly, and was subject to periodic spot checks for quality control.[7]

Despite these measures, uncontrolled tester effects remain a threat to inferences about discrimination. We assess the sensitivity of our results to testers in four ways. First, each tester may have a unique effect, but the average effect of the testers may be zero. In this case, the observations from each tester will be correlated and standard errors that ignore this clustering will tend to be too small. We allow for this possibility by fitting an additional random effect for each tester in our hierarchical logistic regression.[8] Second, each tester may have a unique effect, but these effects may not average to zero. To assess the sensitivity of our results to each tester, we perform a type of cross-validation in which the treatment effect is recalculated for a reduced data set, sequentially omitting those employers associated with each individual tester. Confidence intervals below are based on models that include employer and tester random effects. We compare these results with cross-validation treatment effects based on subsets of the data in which individual testers are sequentially omitted. Third, we recalculate our key results for each unique combination of testers matched in teams over the course of the fieldwork (see Appendix, Table A2). These results, although sensitive to small sample sizes for some combinations, tend to support the consistency of effects across a number of tester comparisons.

As a final investigation of tester effects, we consider the possibility that the expectations or behaviors of testers may influence the audit results in nonrandom ways. For example, if a black tester expects to be treated poorly by

[4] In-person audits also allow for the inclusion of a wide range of entry-level job types (which often require in-person applications); they provide a clear method for signaling race, without concerns over the class-connotations of racially distinctive names (Fryer and Levitt 2004); and they allow us to gather both quantitative and qualitative data, with information on whether an applicant receives the job as well as how he is treated during the interview process.

[5] These 300 applicants were prescreened for appropriate age, race, ethnicity, and gender.

[6] Indeed, as an employer herself, the researcher must identify subtle cues about applicants that indicate their ability to perform. Whether or not these cues are explicit, conscious, or measurable, they are present in a researcher's evaluation of tester candidates, just as they are in employers' evaluations of entry-level job applicants. Like employers, researchers are affected by both objective and subjective/subconscious indicators of applicant quality in their selection and matching of testers in ways that should ultimately improve the nuanced calibration of test partners.

[7] In addition to on-site supervision at the start and finish of each day of fieldwork, on several occasions, we "tested the testers." For example, we hid video cameras in the offices of confederate employers, which allowed us to monitor testers' compliance with the audit protocol as well as to use the tapes as a training tool to better synchronize test partners' performance (not counted among results).

[8] Additional models (not shown here) test for fixed effects of individual testers; we find no significant differences across testers within each race group.

From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

employers, he may appear more withdrawn, nervous, or defensive in interactions. The nature of the interaction may create a self-fulfilling prophecy, in which the tester experiences poor outcomes for reasons unrelated to his race (Steele and Aronsen 1995). We assess these tester effects by analyzing the degree to which personal contact between testers and employers is associated with widening racial disparities. Overall, we find no evidence that testers' interpersonal styles or expectations are associated with increasing discrimination; if anything, personal contact appears to weaken the effect of race, suggesting that testers' performance minimized, rather than exaggerated, our measures of racial bias (see Appendix, Table A1).

The problem of imperfect matching among testers is a well-understood vulnerability of audit experiments, and one to which we devoted considerable attention. Ironically, however, achieving perfect matches can itself produce distortions in the hiring process. Because audit partners are matched on all characteristics that are most directly relevant to hiring decisions (e.g., education, work experience, and physical attributes), employers may be forced to privilege relatively minor characteristics simply out of necessity to break a tie (Heckman 1998:111). If employers care only marginally about race, but are confronted with applicants equal on all other dimensions, this single characteristic may take on greater significance than it would under normal circumstances when evaluating real applicants who differ according to multiple dimensions.

The design of our study, which focuses on the early stages of the hiring process, avoids situations in which employers must choose only a single applicant. By using "callbacks" as one of our key dependent variables, we include cases that represent an employer's first pass at applicant screening.[9] Indeed, recent surveys suggest that employers interview an average of six to eight applicants for each entry-level job opening (Pager 2007b). If race represents only a minor concern for employers, we would expect all members of our audit team to make it through the first cut. If race figures prominently in the first round of review, we can infer that this characteristic has been invoked as more than a mere tie-breaker. In these cases, the evidence of race-based decision making is quite strong.

EXPERIMENTAL RESULTS

The primary results from the field experiment focus on the proportion of applications submitted by testers that elicited either a callback or a job offer from employers, by race of the applicant. Our first team assesses the effects of race discrimination by comparing the outcomes of equally qualified white, Latino, and black applicants. Figure 1a reports positive response rates for each racial/ethnic group. In applications to 171 employers, the white tester received a callback or job offer 31.0 percent of the time, compared with a positive response rate of 25.2 percent for Latinos and 15.2 percent for blacks. These results show a clear racial hierarchy, with whites in the lead, followed by Latinos, and blacks trailing behind.

Figure 1b shows the contrasts between the three race groups. Once we adjust for employer and tester effects, the confidence interval for the white-Latino ratio of 1.23 includes one.[10] By contrast, the white-black ratio of 2.04 is substantively large and statistically significant. The positive response rate for blacks is also significantly lower than the rate for Latinos. The points in the figure show the cross-validation results obtained by sequentially dropping cases associated with each individual tester. All ratios remain consistently greater than one, indicating that employers treat blacks less positively

[9] Positive responses recorded in this study were fairly evenly split between callbacks and job offers. Employers who made offers on-the-spot were typically hiring more than one applicant, thus similarly avoiding a situation in which a forced-choice becomes necessary. In fact, rates of job offers were more evenly distributed by race relative to callbacks (see Tables A1 and A2).

[10] In a model pooling cases from the two teams, with main effects for team and criminal background, the white-Latino gap becomes statistically significant. The generality of this result certainly deserves more study. The Puerto Ricans of New York that our Latino testers represent are a longstanding community of U.S. citizens. In other local labor markets, where markers of citizenship and accent are more prominent sources of difference, evidence of ethnic discrimination may be stronger.

From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

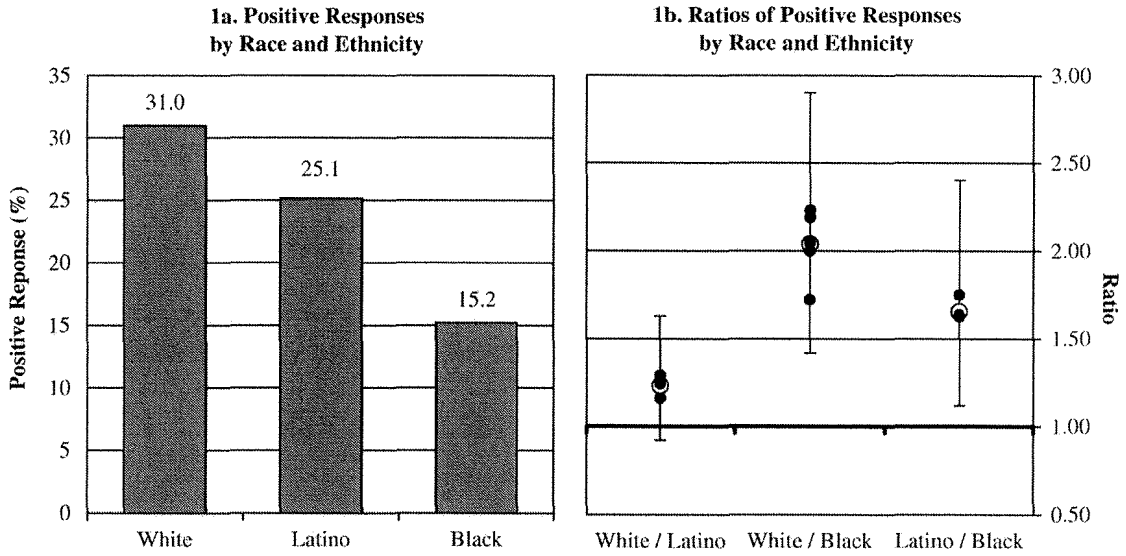


Figure 1. Positive Response Rates and Paired Comparisons by Race and Ethnicity

Notes: Positive responses refer to callbacks or job offers. Hollow circles in Figure 1b indicate point estimates of the ratio. Solid circles indicate ratios obtained by sequentially dropping testers from the analysis. We estimated 95 percent confidence intervals from a hierarchical logistic regression with employer and tester random effects. Number of employers = 171.

regardless of which testers are applying for jobs. Overall, these results indicate that, relative to equally qualified blacks, employers significantly prefer white and Latino job applicants. The findings suggest that a black applicant has to search twice as long as an equally qualified white applicant before receiving a callback or job offer from an employer.

The results from this first comparison indicate employers' strong racial preferences, but the magnitude of this preference remains somewhat abstract. To calibrate the effects of race against another stigmatized category, the ex-offender, we repeated the experiment, this time assigning a criminal record to the white tester. Figure 2a shows the percentage of positive responses—job offers or callbacks—received by each tester. In this experiment, whites with criminal records obtained positive responses in 17.2 percent of 169 job applications, compared with 15.4 percent for Latinos and 13.0 percent for blacks.[11] The white testers' racial advantage narrows substantially in this comparison; yet the white applicant with a criminal record still does just as well, if not better, than his minority counterparts with no criminal background.

Figure 2b shows that the white-Latino ratio is close to one and the confidence interval overlaps one by a large margin. The white-black ratio is now a statistically insignificant 1.32, compared with a significant ratio of 2.04 when the white tester had a clean record. As in the previous experiment, Latinos were preferred to blacks, but here the difference is not significant. As before, the cross-validation treatment effects, obtained by dropping employers associated with one particular tester, are all close to one. These results indicate that, regardless of which testers were sent into the field, employers differentiated little among the three applicant groups.

The comparison of a white felon with black and Latino applicants with clean backgrounds provides a vivid calibration of the effects of race on hiring decisions. While ex-offenders are disadvantaged in the labor market relative to applicants with no criminal background, the stigma of a felony conviction appears to be no greater than that of minority status. Replicating earlier results from Milwaukee (Pager 2003), these findings suggest that New York employ-

[11] The overall rate of positive responses is lower for all testers relative to the results presented in Figure 1. This is likely due to the staggered fielding of teams and resulting differences in the composition of employers audited across the two time periods.

From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

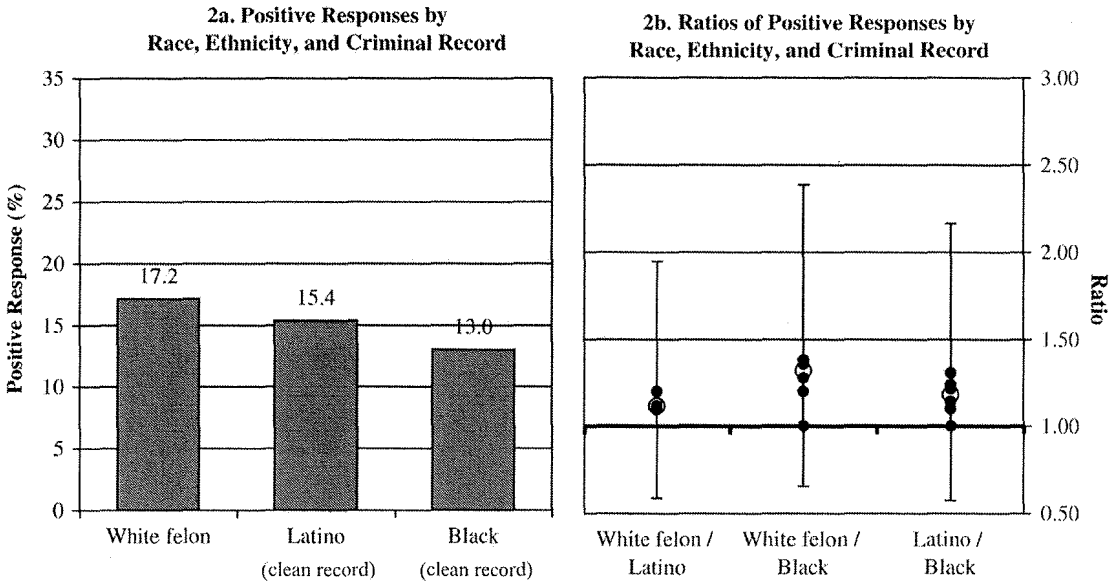


Figure 2. Positive Response Rates and Paired Comparisons by Race, Ethnicity, and Criminal Background

Notes: Positive responses refer to callbacks or job offers. Hollow circles in Figure 2b indicate point estimates of the ratio. Solid circles indicate ratios obtained by sequentially dropping testers from the analysis. We estimated 95 percent confidence intervals from a hierarchical logistic regression with employer and tester random effects. Number of employers = 169.

ers view minority applicants as essentially equivalent to whites just out of prison.

Theories of statistical discrimination point to the very high incarceration rates among young black men as a key explanation for employers' indifference between white felons and blacks with potentially unobserved criminal histories. Current estimates suggest that roughly 18 percent of young black men with high school degrees will experience incarceration by their early 30s (Pettit and Western 2004), and a larger fraction surely have lower level convictions and arrests. Still, the fact that known information about a white applicant's serious criminal conviction is viewed with no more concern than the assumed characteristics of a young black man points to the strength and intensity of contemporary racial attitudes. Overcoming these negative expectations, even for a candidate with otherwise appealing characteristics, requires the negotiation of a number of significant hurdles not present for white job seekers.

Race at Work: An Examination of Interactions between Applicants and Employers

The strong evidence of hiring discrimination from the field experiment provides a clear measure of the continuing significance of race in employer decision making. These numbers, however, tell us little about the process by which race comes to matter. Fortunately, the in-person design of the experiment allows us to further supplement the experimental findings with qualitative evidence from testers' field notes that report their interactions in job interviews. These detailed narratives describe employers' deliberations and suggest some of the ways race comes into play during employment interactions.

Our analysis examines cases in which testers had sufficient interaction with employers for content coding. Consistent with the notion that contemporary forms of discrimination are largely subtle and covert, many cases contain little that would lead us to anticipate the differential treatment that followed. Of those that do, however, we observe several consistent patterns in employers' responses. In particular, three categories of behavior stand out, which we refer to

From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

here as: categorical exclusion, shifting standards, and race-coded job channeling. The first type of behavior, categorical exclusion, is characterized by an immediate or automatic rejection of the black (or minority) candidate in favor of a white applicant. Occurring early in the application process, these decisions involve little negotiated interaction but appear to reflect a fairly rigid application of employers' racial preferences or beliefs. A second category of behavior, shifting standards, reflects a more dynamic process of decision making. Here we observe cases in which employers' evaluations of applicants appear actively shaped or constructed through a racial lens, with similar qualifications or deficits taking on varying relevance depending on an applicant's race. Finally, a third category of behavior moves beyond the hiring decision to a focus on job placement. Race-based job channeling represents a process by which minority applicants are steered toward particular job types, often those characterized by greater physical demands and reduced customer contact.

By observing the interactions that characterize each of these behavior types, we gain a rare glimpse into the processes by which discrimination takes place. At the same time, we emphasize that this discussion is intended as a descriptive exercise rather than a formal causal analysis. Indeed, the categories we identify are not mutually exclusive; some of the same processes may be operating simultaneously, with employers' shifting evaluations of applicant skills leading to different patterns of job channeling, or assumptions about the appropriate race of the incumbent of a particular position leading to forms of categorical exclusion. Likewise, this typology cannot account for all of the differential treatment we observe—at least half of the employer decisions were made on the basis of little or no personal contact between applicant and employer, leaving the nature of the decision entirely unobserved. With these caveats in mind, we nevertheless view the analysis as providing a unique contribution to the study of racial discrimination, revealing mechanisms at work that observational research can rarely identify.

Categorical Exclusion

Few interactions between our testers and employers revealed signs of racial animus or hostility toward minority applicants. At the same time, a close comparison of test partners' experiences shows a number of cases in which race appears to be the sole or primary criterion for an employer's decision. With little negotiation or deliberation over the selection decision, these employers' decisions seem to reflect a preexisting judgment regarding the adequacy or desirability of a minority candidate. The uncompromising nature of the employer's decision can be characterized as a form of categorical exclusion.

A clear-cut case of categorical exclusion was provided when all three testers applied for a warehouse worker position and received a perfunctory decision. Zuri, one of our black testers, reported: "The original woman who had herded us in told us that when we finished filling out the application we could leave because 'there's no interview today, guys!' . . . When I made it across the street to the bus stop . . . the woman who had collected our completed applications pointed in the direction of Simon, Josue, and myself [the three test partners] motioning for us to return. All three of us went over. . . . She looked at me and told me she 'needed to speak to these two' and that I could go back." Zuri returned to the bus stop, while his white and Latino test partners were both asked to come back at 5 p.m. that day to start work. Simon, the white tester, reported, "She said she told the other people that we needed to sign something—that that's why she called us over—so as not to let them know she was hiring us. She seemed pretty concerned with not letting anyone else know."

In this context, with no interview and virtually no direct contact with the employer, we observe a decision that appears to be based on little other than race. The job is a manual position for which Zuri is at least as able, yet he is readily passed over in favor of his white and Latino counterparts.

This case is unusual in that three testers were rarely present at a given location at the same time. More often, we found evidence of differential treatment only after comparing the testers' reports side by side. Here again, we observed several hiring decisions in which race appeared to be the sole or primary source of differentia-

From callbacks to labor markets - (in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.
752

tion. In one example, the three testers inquired about a sales position at a retail clothing store. Joe, one of our black testers, reported that, "[the employer] said the position was just filled and that she would be calling people in for an interview if the person doesn't work out." Josue, his Latino test partner, was told something very similar: "She informed me that the position was already filled, but did not know if the hired employee would work out. She told me to leave my résumé with her." By contrast, Simon, their white test partner, who applied last, had a notably different experience: "I asked what the hiring process was—if they're taking applications now, interviewing, etc. She looked at my application. 'You can start immediately?' Yes. 'Can you start tomorrow?' Yes. '10 a.m.' She was very friendly and introduced me to another woman (white, 28) at the cash register who will be training me."

A similar case arose a few weeks later at an electronics store. Joe, the black tester, was allowed to complete an application but was told that his references would have to be checked before he could be interviewed. Meanwhile, Simon and Josue, his white and Latino partners, applied shortly afterward and were interviewed on the spot. Joe's references were never called, while Simon received a callback two days later offering him the job.

When evaluated individually, these interactions do not indicate racial prejudice or discrimination. Side by side, however, we see that minority applicants encounter barriers not present for the white applicant, with employers citing excuses for putting off the black or minority candidate (e.g., "the job has already been filled" or "we'd have to check your references before we can proceed") that appear not to apply for the white applicant. To be sure, certain cases may capture random error—perhaps a position became available between the testers' visits, or an employer was otherwise preoccupied when one applicant arrived but not another, leading to the employer's differential response. Still, the consistency of the pattern in these data suggests that random error is unlikely to be a dominant factor. Indeed, of the 171 tests conducted by the first team (no criminal background), white testers were singled out for callbacks or job offers 15 times, whereas there was only a single case in which a black tester received a positive response when his white or Latino partner did not.[12]

These cases of categorical exclusion, although directly observed in only a small number of audits (5 of the 47 cases of differential treatment across the two teams), reveal one form of discrimination in which racial considerations appear relatively fixed and unyielding.[13] Before black (or minority) candidates have the chance to demonstrate their qualifications, they are weeded out on the basis of a single categorical distinction.

Categorical exclusion represents one important form of discrimination. While these rather abrupt interactions reveal little about the underlying motivation that drives employers' decisions, they do demonstrate the sometimes rigid barriers facing minority job seekers. In these cases, black (or minority) applicants are discouraged or dismissed at the outset of the employment process, leaving little opportunity for a more nuanced review.

Shifting Standards

Making it past the initial point of contact was not the only hurdle facing minority applicants. Indeed, among those who recorded more extensive interaction with employers, we observe a complex set of racial dynamics at work. On the one hand, personal contact with employers was associated with significantly improved outcomes for all testers and a narrowing of the racial gap (see Appendix, Table A1). The testers' interpersonal skills seemed to reduce the influ-

[12] In an additional 13 cases, both white and Latino testers received positive responses; in seven cases, the Latino tester alone was selected (see Appendix, Table A2).

[13] The denominator of 47 represents the total number of cases of black-white differential treatment from the first (N = 28) and second (N = 19) teams. In calculating the numerator, we do not include a number of additional cases of differential treatment resulting from applications in which there was little or no personal contact between testers and the employer (rates of personal contact were similar by race of tester). In such cases, differential treatment may reflect categorical exclusion (based on a visual assessment of the candidate), shifting standards (based on a review of the completed applications), random error, or something else.

From callbacks to labor markets - In-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

ence of racial bias, or at least did not exacerbate it. Yet, even in the context of this more personalized review, we see evidence of subtle bias in the evaluation of applicant qualifications. In particular, a number of cases reveal how testers' "objective" qualifications appear to be reinterpreted through the lens of race. Although testers' résumés were matched on education and work experience, some employers seemed to weigh qualifications differently depending on the applicant's race. In the following interactions, we see evidence that the same deficiencies of skill or experience appear to be more disqualifying for the minority job seekers (N = 11).

In one case, Joe, a black tester, was not allowed to apply for a sales position due to his lack of direct experience. He reported, "[The employer] handed me back my résumé and told me they didn't have any positions to offer me . . . that I needed a couple years of experience." The employer voiced similar concerns with Josue and Kevin, Joe's Latino and white partners. Josue wrote, "After a few minutes of waiting . . . I met with [the employer] who looked over my résumé. He said that he was a little worried that I would not be able to do the work." Kevin reported an even stronger reaction: "[The employer] looked at my résumé and said, 'There is absolutely nothing here that qualifies you for this position.'" Yet, despite their evident lack of qualifications, Kevin and Josue were offered the sales job and asked to come back the next morning. In interactions with all three testers, the employer clearly expressed his concern over the applicants' lack of relevant work experience. This lack of experience was not grounds for disqualification for the white and Latino candidates, whereas the black applicant was readily dismissed.

When applying for a job as a line cook at a midlevel Manhattan restaurant, the three testers encountered similar concerns about their lack of relevant experience. Josue, the Latino tester, reported, "[The employer] then asked me if I had any prior kitchen or cooking experience. I told him that I did not really have any, but that I worked alongside cooks at [my prior job as a server]. He then asked me if I had any 'knife' experience and I told him no. . . . He told me he would give me a try and wanted to know if I was available this coming Sunday at 2 p.m." Simon, his white test partner, was also invited to come back for a trial period. By contrast, Joe, the black tester, found that "they are only looking for experienced line cooks." Joe wrote, "I started to try and convince him to give me a chance but he cut me off and said I didn't qualify." None of the testers had direct experience with kitchen work, but the white and Latino applicants were viewed as viable prospects while the black applicant was rejected because he lacked experience.

In other cases, employers perceived real skill or experience differences among applicants despite the fact that the testers' résumés were designed to convey identical qualifications. In one example, the testers applied for a job at a moving company. Joe, the black applicant, spoke with the employer about his prior experience as a stockperson at a moving truck company, but "[the employer] told me that he couldn't use me because he is looking for someone with moving experience." Josue, his Latino partner, presented his experience as a stocker at a delivery company and reported a similar reaction, "He then told me that since I have no experience . . . there is nothing he could do for me." Simon, their white test partner, presented identical qualifications, but the employer responded more favorably: "'To be honest, we're looking for someone with specific moving experience. But because you've worked for [a storage company], that has a little to do with moving.' He wanted me to come in tomorrow between 10 and 11." The employer is consistent in his preference for workers with relevant prior experience, but he is willing to apply a more flexible, inclusive standard in evaluating the experience of the white candidate than in the case of the minority applicants. Employers' shifting standards, offering more latitude to marginally skilled white applicants than to similarly qualified minorities, suggest that even the evaluation of "objective" information can be affected by underlying racial considerations.

Even in cases where the white tester presented as a felon, we see some evidence that this applicant was afforded the benefit of the doubt in ways that his minority counterparts were not. In applying at an auto dealership, for example, the three testers met with very different reactions. Joe, the black tester, was informed at the outset that the only available positions were for people with direct auto sales experience. When Josue, his Latino partner, applied, the lack of direct auto sales experience was less of a prob-

From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.
754

lem. Josue reported, "He asked me if I had any customer service experience and I said not really. . . . He then told me that he wanted to get rid of a few bad apples who were not performing well. He asked me when I could start." Josue was told to wait for a callback on Monday. When the employer interviewed Keith, their white ex-felon test partner, he gave him a stern lecture regarding his criminal background. The employer warned, "I have no problem with your conviction, it doesn't bother me. But if I find out money is missing or you're not clean or not showing up on time I have no problem ending the relationship." Despite the employer's concerns, Keith was offered the job on the spot. The benefit of the doubt conferred by whiteness persists here, even in the context of a white applicant just released from prison.

A pattern in these interactions, when compared side by side, is the use of double standards—seeking higher qualifications from blacks than non-blacks, or viewing whites as more qualified than minorities who present equivalent résumés. Recent research emphasizes employers' use of race as a proxy for difficult-to-observe productivity characteristics (Moss and Tilly 2001; Waldinger and Lichter 2003). Where we have detailed field notes on job interviews, the interactions we observe suggest that employers also use race in interpreting and weighing observable skill characteristics. Standards appear to shift as employers evaluate various applicants' qualifications differently depending on their race or ethnicity (see also Biernat and Kobrynowicz 1997; Yarkin, Town, and Wallston 1982).

Race-Coded Job Channeling

The first two categories of differential treatment focus on the decision to hire. Beyond this binary decision, employers also face decisions about where to place a worker within the organizational hierarchy. Here, at the point of job placement, we observe a third category of differential treatment. In our review of the testers' experiences, we noticed that applicants were sometimes encouraged to apply for different jobs than the ones initially advertised or about which they had inquired. In many cases, these instances of channeling suggest a race-coding of job types, whereby employers prefer whites for certain positions and minorities for others. In one case, Zuri, a black tester, applied for a sales position at a lighting store that had a sign in the front window stating "Salesperson Wanted." Zuri described the following interaction: "When she asked what position I was looking for I said I was open, but that since they were looking for a salesperson I would be interested in that. She smiled, put her head in her hand and her elbow on the table and said, 'I need a stock boy. Can you do stock boy?'" Zuri's white and Latino test partners, by contrast, were each able to apply for the advertised sales position.

Another black tester, Joe, was similarly channeled out of a customer service position in his application to a Japanese restaurant. Joe reported, "I told her I was there to apply for the waiter position and she told me that there were no server positions. I told her it was advertised in the paper, and she said there must have been a mistake. She said all she had available was a busboy position. I told her since there was no waiter position, I would apply for the busboy." Later that day, Kevin, his white test partner, was hired for the server position on the spot.

We also observed channeling of the Latino testers. Josue's fieldnotes of an audit at a clothing retailer begin by describing the "young white 20-something women running the place." One of the women interviewed him, asked about past work experience, and asked which job he was applying for. "I told her 'sales associate,'" Josue reported, and he presented a résumé on which the most recent job listed was as a sales assistant at a sporting goods store. "She then told me that there was a stock position and asked if I would be interested in that." Josue was offered the stocker job and asked to start the next day.

In many cases, these instances of channeling are coded as "positive responses" in the initial analyses. While our key concern is about access to employment of any kind, this general focus masks another form of racial bias at work. A closer analysis of the testers' experiences suggests that decisions about job placement, like hiring more generally, often follow a racial logic. We coded all instances of job channeling across both our teams and counted 53 cases (compared with 172 positive responses). By comparing the original job title to the suggested job type, we then categorized these cases as downward channeling, upward channeling, lateral channeling, or unknown. We define downward channeling as (1) a move from a job

From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

involving contact with customers to a job without (e.g., from server to busboy), (2) a move from a white-collar position to a manual position (e.g., from sales to stocker), or (3) a move in which hierarchy is clear (e.g., from supervisor to line worker). We define upward channeling as a move in the opposite direction. We focus on these two types of channeling for our current analysis. After eliminating cases in which all testers within a team were similarly channeled, we have 23 additional cases of differential treatment that were not recorded by our initial measurement of job offers and callbacks.

Like hiring criteria, job placement is also patterned by race (see Table 1). Black applicants were channeled into lower positions in nine cases, Latinos were channeled down in five cases, and whites experienced downward channeling in only one case. Many of these cases were restaurant jobs in which the tester applied for a position as a server but was steered to a job as a busboy or dishwasher. In almost all cases, the original position required extensive customer contact while the suggested position did not (e.g., salesperson to stocker). Testers were sometimes guided into lower positions because their résumés indicated limited work experience, but racial differences in channeling suggest that insufficient work experience was more penalizing for minorities than for whites. The one case of downward channeling among white applicants involved a tester presenting with a criminal background.

In fact, whites were more often channeled up than down. In at least six cases, white testers were encouraged to apply for jobs that were of a higher-level or required more customer contact than the initial position they inquired about. In one case, the white tester was even encouraged to apply for a supervisory position, despite limited work experience. Kevin reported: "[The employer] then asked me if I had any experience in construction. I told him I did not. He asked if I would be okay working with people that have thick accents like his. I told him that was fine. He then told me that he wanted me to be his new company supervisor."

Employers appear to have strong views about what kind of person is appropriate for what kind of job, based either on their own assumptions of worker competence or assumptions about what their clients expect or prefer in the appearance of those serving them. Consistent with the testers' field notes, employers appear to apply more stringent hiring criteria to minority workers, preferring whites for jobs that require greater skill or responsibility. In addition, minorities are disproportionately channeled out of customer service positions, consistent with other research in which employers view minority applicants as lacking communication skills or as otherwise discomfiting for customers. Although our testers presented highly effective styles of interpersonal communication, the cursory review process for these jobs often leaves group membership more salient than any individuating characteristics.

Table 1. Job Channeling by Race

Original Job Title	Suggested Job
Blacks Channeled Down	
Server	Busser
Counter person	Dishwasher/porter
Server	Busboy
Assistant manager	Entry fast-food position
Server	Busboy/runner
Retail sales	Maintenance
Counter person	Delivery
Sales	Stockboy
Sales	Not specified[a]
Latinos Channeled Down	
Server	Runner
Sales	Stock
Steam cleaning	Exterminator
Counter person	Delivery
Sales	Stock person
Whites Channeled Down	
Server	Busboy
Latinos Channeled Up	
Carwash attendant	Manager
Warehouse worker	Computer/office
Whites Channeled Up	
Line cook	Waitstaff
Mover	Office/telesales
Dishwasher	Waitstaff
Driver	Auto detailing
Kitchen job	"Front of the house" job
Receptionist	Company supervisor

Note: This table includes all cases of upward and downward channeling, except when all testers on a team were channeled similarly.
[a] Employer told tester that "sales might not be right for you."

From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.
756

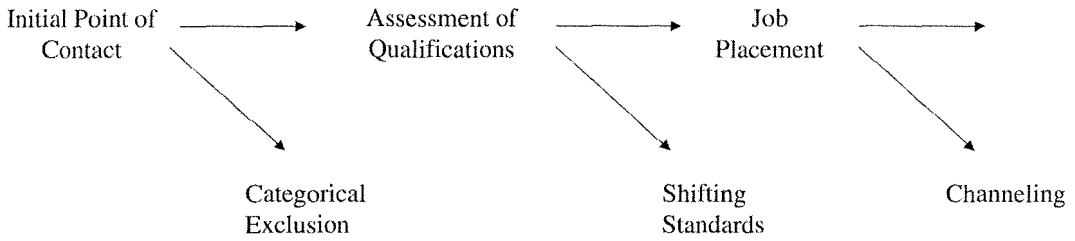


Figure 3. Discrimination at Three Decision Points

The three types of differential treatment we observe illustrate how employers enact their racial preferences in the hiring process. Rather than outward hostility or racial animus, we see more subtle forms of discouragement or rejection. At multiple points in the hiring process, black (or Latino) applicants face additional hurdles or barriers that reduce their chances of employment and affect the quality of jobs for which they are considered. Figure 3 illustrates the processes identified in the preceding discussion. At each of the three decision points, we see pathways deflected by various forms of racial bias. Subtle differences in employers' responses—often imperceptible to the applicants themselves—produce a pattern of outcomes systematically affected by race.

Complementing the quantitative indicators of differential treatment, these qualitative observations provide a rare window into the processes by which discrimination occurs. The three categories of differential treatment observed in these data point to the range of experiences that constitute discrimination in the employment process.[14] In a small number of cases, minority testers were disqualified early on in decisions that appear to reflect employers' fairly rigid preferences. These instances of categorical exclusion represent one of the most extreme forms of discrimination, wherein minority applicants have little opportunity to overcome employers' potential concerns. By contrast, a larger number of interactions suggest a more complicated set of negotiations at play. In evaluating applicant qualifications, minority applicants, and black men in particular, appear to be held to a higher standard than their white counterparts. Black men are disqualified more readily, or hired more reluctantly, than their white partners with identical skills and experience. Furthermore, racialized assessments of applicant quality and "fit" affect not only the decision to hire, but also decisions about job placement, with minority applicants more often channeled into positions involving less skill or customer contact. Together, these experiences illustrate how racial disadvantage is dynamically constructed and reinforced, with the assessment of applicant qualifications and suitability subject to interpretation and bias. While not an exhaustive catalogue of discrimination experiences, the fact that these dynamics are observed in natural settings (with little prompting) attests to their relative frequency and regularity. Our testers' experiences suggest how race shapes employers' evaluations in subtle but systematic ways, with important implications for structuring opportunity along racial lines.

DISCUSSION

Sending trained testers with equivalent résumés to apply for entry-level jobs reveals clear evidence of discrimination among low-wage employers in New York City. Blacks were only half as likely to receive a callback or job offer relative to equally qualified whites; moreover, black and Latino applicants with clean backgrounds fared no better than a white applicant

[14] To be sure, our study captures only a few of the many pathways in the employment process that are potentially affected by racial bias. Beyond our window of observation, the pathways of this diagram would presumably continue along later points in the employment process, including wage-setting decisions, training opportunities, promotion, and termination decisions. This research represents one incremental contribution to understanding—and documenting—the varied decision points that may be affected by race.

From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

just released from prison. The magnitude of these racial disparities provides vivid evidence of the continuing significance of race in contemporary low-wage labor markets. There is a racial hierarchy among young men favoring whites, then Latinos, and finally blacks as the candidates of last resort.

The episodes of discrimination recorded in this study were seldom characterized by overt racism or hostility. In fact, our testers rarely perceived any signs of clear prejudice. It was only through side-by-side comparisons of our testers' experiences that patterns of subtle but consistent differential treatment were revealed. Minority applicants were disqualified more readily and hired more reluctantly than their white partners with identical skills and experience. Additionally, black and Latino applicants were routinely channeled into positions requiring less customer contact and more manual work than their white counterparts. In interactions between applicants and employers, we see a small number of cases that reflect employers' seemingly rigid racial preferences. More often, differential treatment emerged in the social interaction of the job interview. Employers appeared to see more potential in the stated qualifications of white applicants, and they more commonly viewed white applicants as a better fit for more desirable jobs.

Our findings of discrimination are particularly striking because the testers in this study represent a best-case scenario for low-wage job seekers. The testers were college-educated young men with effective styles of self-presentation. Although posing as high school graduates with more limited skills, these young men stood well above the typical applicant for these low-wage jobs. The effects of race among individuals with fewer hard and soft skill advantages may well be larger than those estimated here.

At the same time, while we find robust evidence of racial discrimination, we should be careful not to interpret these results as showing the level of discrimination actively experienced by minority job seekers in the New York labor market. Our sampling design, based on employers, not workers, over-represents small firms relative to their share of employment. The sample includes many restaurants and independent retailers for whom hiring is less bureaucratic, and who lack the human resource departments that manage the equal employment opportunity obligations of large firms (Dobbin et al. 1993). Nevertheless, our sampled employers well represent the kinds of low-skill service work that dominate low-wage urban labor markets.

A second limitation on the generalizability of our findings results from our sampling procedures based on classified advertisements. Surveys of job seekers suggest that 25 to 30 percent of low skill jobs are filled by classified ads; the remainder are filled through some combination of network referrals, walk-in applications, and employment agencies (Holzer 1987). These search strategies may generate a different distribution of employers from that reported here. Some argue that the focus on jobs advertised through metropolitan newspapers understates the extent of discrimination. Firms that wish to discriminate, it is argued, are more likely to advertise job openings through more restrictive channels, such as networks of existing employees, employment agencies, or more selective publications (Elliott 2000; Fix and Struyk 1993; Petersen, Saporta, and Seidel 2000). Others, by contrast, argue that any random sample of employers will overstate the extent of discrimination actually experienced by job seekers. If minority applicants can identify and avoid firms that discriminate, the actual incidence of labor market discrimination will be correspondingly reduced (Becker 1957; Heckman 1998).

Of course, minority workers' ability to avoid the effects of discrimination by self-selecting into nondiscriminatory firms requires that a sufficient number of nondiscriminatory employers exist; that there are no differences in the quality of jobs offered by employers who are more and less likely to discriminate; and that the search costs necessary to locate nondiscriminatory employers are trivial. Future research using microdata to track the search patterns and outcomes of black and white job seekers could better address these issues. From our data, we can safely conclude that job searches across a wide range of employers represented by the classified ads of five New York newspapers reveal substantial discrimination. Understanding how job seekers adapt to this reality remains a challenge for future research.

Our findings for the New York City labor market add to evidence of racial discrimination in employment reported from recent field

From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.
758

experiments in Milwaukee, Boston, and Chicago (Bertrand and Mullainathan 2004; Pager 2003). The significant evidence of discrimination found in these studies contrasts sharply with recent survey research showing small racial differences in wages (Farkas and Vicknair 1996; Neal and Johnson 1996). How might these disparate findings be reconciled? First, as noted above, the presence of discrimination in the labor market may lead workers to differentially sort across employers, such that minority job seekers queue for jobs offered by employers who are less likely to discriminate. These dynamics can lead to longer search or wait times for minority job seekers, which might not be reflected in ultimate wage offers. Indeed, data from the late 1990s show that the unemployment spells of black men (3.1 months) are about twice as long as those for whites (1.6 months) (Gottschalck 2003:2). This suggests that the primary effects of discrimination on labor market outcomes may be reflected in employment differentials rather than wages.[15]

Second, the experience of discrimination may add to the psychic costs of the job search process, prompting some to opt out altogether. If discrimination discourages all but the most motivated and able black job seekers, black wage earners would represent an increasingly select group. Since the 1990s, increasing numbers of young black men have dropped out of the formal labor market, contributing to an artificial convergence of black and white wages (Western and Pettit 2005). Without effectively accounting for the processes that precede labor force participation—including the discouraging effects of discrimination—wage estimates can account for only one incomplete picture of the larger employment process.

Our findings add to a large research program demonstrating the continuing contribution of discrimination to racial inequality in the post-civil rights era. Still, significant questions remain unanswered. The audit experiment necessarily focuses on employers' hiring behaviors but does not examine the skills, preferences, and networks of job seekers. We do not know, and few research designs have been devised to test, the relative magnitude of the effects of discrimination compared with the effects of human and social capital. Such an analysis would need to study both employers as they screen job applicants and workers as they search for jobs.

The effects of discrimination, relative to human and social capital, should also be defined broadly. As evidence of discrimination in the post-civil rights era accumulates, new research should go beyond determining whether discrimination is present to consider how the effects of discrimination unfold over the life cycle and across social space. Episodes of discrimination may not only cause unemployment at one point in time, but may have long-term effects that weaken minority workers' attachment to the labor market and reduce labor force participation. Discrimination may produce broad cultural effects in which work itself is de-legitimated as a fair source of opportunity. The effects of discrimination may also vary across the population, concentrating perhaps among the young men whose employment rates are lowest. Tracing these larger and more varied effects of discrimination show both the advantages and limits of the experimental method used here. The experiment allows us to infer discrimination with great certainty, but the effects of discrimination are narrowly defined. The broader effects of discrimination—on the cultural dimensions of economic life and over the life course—are harder to pinpoint but may indicate more fundamental and intractable inequalities. A research agenda that includes these wider consequences would be less skeptical that discrimination exists and more curious about its continuing effects on not just employment inequality, but on American race relations more broadly.

[15] Johnson and Neal (1998), for example, find that after controlling for cognitive ability and other human capital characteristics, black-white differences in employment among young men remain large and statistically significant. The importance of employment over wages for racial inequality in economic status is likely to be especially great for young non-college-educated men, for whom the overall level of wage dispersion is low. Later in the life course, as wage dispersion increases and labor force experience accumulates, the racial wage gap becomes more pronounced (Tomaskovic-Devey, Thomas, and Johnson 2005). For a historical example, see Whatley (1990), who shows that despite the substantial racial barriers to employment that existed among Northern firms after World War I, blacks and whites experienced remarkably similar wage rates.

From callbacks to labor markets - In-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.
759

Devah Pager *is an Associate Professor of Sociology and Faculty Associate of the Office of Population Research at Princeton University. Her book,* Marked: Race, Crime, and Finding Work in an Era of Mass Incarceration, *investigates the racial and economic consequences of large scale imprisonment for contemporary U.S. labor markets.*

Bruce Western *is Professor of Sociology and director of the Multidisciplinary Program in Inequality and Social Policy at Harvard University.*

Bart Bonikowski *is a PhD candidate in sociology at Princeton University. In addition to his work on stratification, he is completing a project that compares popular conceptions of the nation among the populations of 30 countries.*

APPENDIX

Robustness Checks

We examine the robustness of our primary results by looking at racial and ethnic contrasts for different subsets of the data (Table A1). Although small numbers in certain cells lead to some instability in estimates, these breakdowns can examine the consistency of effects across the full range of the sample. To account for learning or adaptation by the testers, we estimate effects for the first and second halves of the experimental period. In each period, whites and Latinos received significantly more positive responses than did blacks, and whites received slightly more positive responses than did Latinos. To examine whether our results depend strongly on any particular area within New York, we separate the experimental effects by location. Over half the audited employers were located in Manhattan. We found the pattern of black disadvantage throughout Manhattan and in the outer boroughs. To examine whether the first tester sent to an employer was more likely to be successful, we randomized the order in which testers were sent. Experimental effects are similar regardless of which tester interviewed first. Finally, we compare the outcomes of audits in which testers had little or no interaction with the employer with those characterized by more substantial personal contact. Here we see some evidence that personal contact reduces racial disparities in employment, consistent with the notion that individualizing information can help offset the effects of negative stereotypes.

The bottom half of Table A1 presents these same comparisons for the team in which the white applicants presented evidence of a felony conviction. Across these comparisons, we find treatment effects close to zero, supporting the finding that employers did not distinguish strongly between whites with criminal records and minority job seekers without. In short, these results indicate a large racial preference among New York employers for white job applicants over black applicants, smaller preference for whites over Latinos and Latinos over blacks, and little difference between white felons and minorities with clean backgrounds. All results are robust to tester effects and experimental effects and appear to be roughly uniform across New York City.

Results by Tester Teams

In the course of fielding two three-person teams of testers, we used 10 different testers: two Latinos, four blacks, and four whites. In each three-person team consisting of a white, a black, and a Latino, we combined the 10 testers into six different unique combinations. Before pooling the data across combinations of testers, Heckman and Siegelman (1993) recommend testing for the homogeneity of responses across combinations. The columns in Table A2 represent mutually exclusive outcomes; overall response rates by race can be calculated by summing all columns in which a given race group is represented. A chi-square test within each team fails to reject the null hypothesis of homogeneity across combinations. With this evidence of homogeneity, we report treatment effects pooled across testers. Table A2 reports the detailed experimental results for each unique combination of testers.

From callbacks to labor markets - In-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.
760

Table A1. Percentage of Positive Responses and Race Differences, by Date, Employer Address, and Race of First Tester

Subsample (N)	White (W)	Latino (L)	Black (B)	Race Differences[a] W/L	W/B	L/B
Total (171)	31.0	25.1	15.2	1.2 (.02)	2.0 (.00)	1.7 (.00)
Date[b]						
Feb. 23 to April 7 (84)	29.8	23.8	9.5	1.3 (.08)	3.1 (.00)	2.5 (.00)
April 8 to July 16 (84)	33.3	27.4	21.4	1.2 (.04)	1.6 (.00)	1.3 (.05)
Location[c]						
Below 34th St. (56)	23.2	21.4	12.5	1.1 (.31)	1.9 (.00)	1.7 (.03)
34th St. to 72nd St. (46)	30.4	21.7	17.4	1.4 (.02)	1.8 (.00)	1.3 (.15)
Above 72nd St. (18)	33.3	22.2	5.6	1.5 (.00)	6.0 (.00)	4.0 (.00)
Other (50)	40.0	34.0	20.0	1.2 (.12)	2.0 (.00)	1.7 (.00)
Race of First Tester						
White (68)	27.9	23.5	10.3	1.2 (.11)	2.7 (.00)	2.3 (.00)
Black (45)	40.0	31.1	20.0	1.3 (.06)	2.0 (.00)	1.6 (.01)
Latino (53)	28.3	22.6	18.9	1.3 (.09)	1.5 (.00)	1.2 (.15)
Type of Positive Response[d]						
Callback (171)	12.9	9.9	2.9	1.3 (.10)	4.4 (.00)	3.4 (.00)
Job offer (171)	21.1	17.0	12.9	1.2 (.02)	1.6 (.00)	1.3 (.02)
Personal Contact[e]						
No personal contact (46)[f]	10.9	6.5	0	1.7 (.09)	>10.9	>6.5
Personal contact (65)	52.3	46.2	29.2	1.1 (.11)	1.8 (.00)	1.6 (.00)

Subsample (N)	White felon (Wf)	Latino (L)	Black (B)	Race Differences[a] Wf/L	Wf/B	L/B
Total (169)	17.2	15.4	13.0	1.1 (.25)	1.3 (.08)	1.2 (.17)
Date[b]						
March 2 to April 13 (83)	16.9	13.3	10.8	1.3 (.16)	1.6 (.06)	1.2 (.21)
April 14 to Aug. 6 (82)	17.1	17.1	15.9	1.0 (.43)	1.1 (.35)	1.1 (.34)
Location[c]						
Below 34th St. (51)	9.8	7.8	3.9	1.3 (.30)	2.5 (.05)	2.0 (.00)
34th St. to 72nd St. (46)	13.0	17.4	13.0	.8 (.74)	1.0 (.42)	1.3 (.14)
Above 72nd St. (7)	0	0	0			
Other (62)	29.0	21.0	21.0	1.4 (.08)	1.4 (.09)	1.0 (.46)
Race of First Tester						
White (53)	20.8	18.9	13.2	1.1 (.34)	1.6 (.13)	1.4 (.15)
Black (59)	18.6	15.3	15.3	1.2 (.20)	1.2 (.15)	1.0 (.39)
Latino (52)	11.5	11.5	11.5	1.0 (.44)	1.0 (.42)	1.0 (.41)
Type of Positive Response[d]						
Callback (169)	11.2	9.5	5.3	1.2 (.23)	2.1 (.01)	1.8 (.02)
Job offer (169)	5.9	6.5	7.7	.9 (.58)	.8 (.77)	.8 (.65)
Personal Contact[e]						
No personal contact (75)	8.0	8.0	4.0	1.0 (.45)	2.0 (.09)	2.0 (.04)
Personal contact (39)	35.9	28.2	30.8	1.3 (.12)	1.2 (.24)	.9 (.58)

[a] Numbers in parentheses are bootstrap *p*-values for a one-sided test of whether the ratio is less than or equal to one.

[b] Changes over time capture several possible effects: learning or adaptation by testers, compositional changes in the types of employers brought into the sample at different points, and changes in the business cycle.

[c] Street addresses are for Manhattan.

[d] Because some testers received both a job offer and a subsequent callback, the sum of these two columns may be greater than the total listed above (in which a positive response is calculated by the presence of a callback *or* a job offer).

[e] Analyses of "personal contact" include only those cases in which all tester partners experienced personal contact; the "no personal contact" analyses include cases in which none of the testers experienced personal contact. This exclusion avoids any confounding effect of employers' racial preferences as reflected in the decision to interview.

[f] Because the response rate for blacks in this subsample is zero, ratios in which blacks are in the denominator are undefined. For the purposes of this analysis, we represent this ratio as greater than the value of the numerator over one.

From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.
761

Table A2. Detailed Experimental Results, by Unique Combination of Testers

	Who Gets a Positive Response (percent)								
Group	All	None	W + L	W + B	L + B	W	L	B	N
White without criminal record (posterior predictive probability of χ^2 statistic: .054)[a]									
1	11	69.2	4.4	3.3	0	7.7	4.4	0	91
2	7.5	67.9	11.3	0	0	9.4	3.8	0	53
3	36.4	18.2	0	0	0	18.2	18.2	9.1	11
4	33.3	33.3	33.3	0	0	0	0	0	6
5	28.6	57.1	14.3	0	0	0	0	0	7
6	0	66.7	0	0	0	33.3	0	0	3
Total	12.9	63.7	7.6	1.8	0	8.8	4.7	.6	171
White with criminal record (posterior predictive probability of χ^2 statistic: .588)									
1	3.7	75.3	2.5	2.5	1.2	7.4	4.9	2.5	81
2	4.9	56.1	2.4	2.4	7.3	14.6	7.3	4.9	41
3	2.8	77.8	8.3	2.8	2.8	2.8	0	2.8	36
4	0	60	0	0	20	0	20	0	5
5	0	75	0	0	0	0	0	25	4
6	0	100	0	0	0	0	0	0	2
Total	3.6	71.0	3.6	2.4	3.6	7.7	4.7	3.6	169

Note: W = white; L = Latino; B = black. Columns of "Who Gets a Positive Response" represent mutually exclusive categories (i.e., rows sum to 100 percent). In the first experiment (no criminal record), there was only a single case (group 3) in which a black tester received a callback when neither of his test partners received one.

[a] The chi-square test is undefined with marginal counts of zero. We calculate a posterior predictive p-value by simulating counts under independence for nonzero cells.

REFERENCES

Agresti, Alan. 1990. *Categorical Data Analysis.* New York: Wiley.

Aigner, Dennis J. and Glen G. Cain. 1977. "Statistical Theories of Discrimination in Labor Market." *Industrial and Labor Relations Review* 30:749–76.

Altonji, Joseph G. and Charles R. Pierret. 2001. "Employer Learning and Statistical Discrimination." *Journal of Economics* 116:313–50.

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

Bendick, Marc, Jr., Charles Jackson, and Victor Reinoso. 1994. "Measuring Employment Discrimination through Controlled Experiments." *Review of Black Political Economy* 23:25–48.

Bendick, Marc, Jr., Charles Jackson, Victor Reinoso, and Laura Hodges. 1991. "Discrimination against Latino Job Applicants: A Controlled Experiment." *Human Resource Management* 30:469–84.

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* 94:991–1013.

Biernat, Monica and Diane D. Kobrynowicz. 1997. "Gender and Race-Based Standards of Competence: Lower Minimum Standards but Higher Ability Standards for Devalued Groups." *Journal of Personality and Social Psychology* 72:544–57.

Bodenhausen, Galen. 1988. "Stereotypic Biases in Social Decision Making and Memory: Testing Process Models of Stereotype Use." *Journal of Personality and Social Psychology* 55(5):726–37.

Cancio, A. Silvia, T. David Evans, and David J. Maume. 1996. "Reconsidering the Declining Significance of Race: Racial Differences in Early Career Wages." *American Sociological Review* 61:541–56.

Cross, Harry, Genevieve Kenney, Jane Mell, and Wendy Zimmerman. 1990. *Employer Hiring Practices: Differential Treatment of Hispanic and Anglo Job Seekers.* Washington, DC: Urban Institute Press.

Darity, William, Jr. and Patrick A. Mason. 1998. "Evidence on Discrimination in Employment: Codes of Color, Codes of Gender." *The Journal of Economic Perspectives* 12:63–90.

Darity, William, Jr. and Samuel L. Meyers. 1998. *Persistent Disparity: Race and Economic Inequality in the United States since 1945.* Cheltenham, UK: Edward Elgar.

Darley, John M. and Paget H. Gross. 1983. "A Hypothesis-Confirming Bias in Labeling Effects." *Journal of Personality and Social Psychology* 44:20–33.

Dobbin, Frank, J. Sutton, J. Meyer, and W. R. Scott. 1993. "Equal Opportunity Law and the

From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Construction of Internal Labor Markets." *American Journal of Sociology* 99:396–427.

Dovidio, John F. and Samuel L. Gaertner. 2000. "Aversive Racism and Selection Decisions." *Psychological Science* 11:315–19.

Elliott, James R. 2000. "Class, Race, and Job Matching in Contemporary Urban Labor Markets." *Social Science Quarterly* 81:1036–52.

Farkas, George and Keven Vicknair. 1996. "Appropriate Tests of Racial Wage Discrimination Require Controls for Cognitive Skill: Comment on Cancio, Evans, and Maume." *American Sociological Review* 61:557–60.

Fiske, Susan. 1998. "Stereotyping, Prejudice, and Discrimination." Pp. 357–411 in *The Handbook of Social Psychology,* edited by D. Gilbert, S. Fiske, and G. Lindzey. Boston, MA: McGraw Hill.

Fix, Michael and Raymond J. Struyk, eds. 1993. *Clear and Convincing Evidence: Measurement of Discrimination in America.* Washington, DC: Urban Institute Press.

Fryer, Ronald G., Jr. and Steven D. Levitt. 2003. "The Causes and Consequences of Distinctively Black Names." NBER Working Paper 9938.

———. 2004. "The Causes and Consequences of Distinctively Black Names." *The Quarterly Journal of Economics* 119:767–805.

Ghosh, Malay, Ming-Hui Chen, Atalanta Ghosh, and Alan Agresti. 2000. "Hierarchical Bayesian Analysis of Binary Matched Pairs Data." *Statistica Sinica* 10:647–57.

Gilbert, Daniel T. and J. Gregory Hixon. 1991. "The Trouble of Thinking: Activation and Application of Stereotypical Beliefs." *Journal of Personality and Social Psychology* 60(4):509–17.

Gottschalck, Alfred O. 2003. "Dynamics of Economic Well-Being: Spells of Unemployment, 1996–1999." *Current Population Reports.* Washington, DC: US Census Bureau.

Heckman, James J. 1998. "Detecting Discrimination." *The Journal of Economic Perspectives* 12:101–116.

Heckman, James and Peter Siegelman. 1993. "The Urban Institute Audit Studies: Their Methods and Findings." Pp. 187–258 in *Clear and Convincing Evidence: Measurement of Discrimination in America,* edited by M. Fix and R. Struyk. Lanham, MD: University Press of America.

Holzer, Harry J. 1987. "Informal Job Search and Black Youth Unemployment." *American Economic Review* 77:446–52.

———. 1996. *What Employers Want: Job Prospects of Less-Educated Workers.* New York: Russell Sage Foundation.

Holzer, Harry J. and Paul Offner. 2001. "Trends in Employment Outcomes of Young Black Men." Georgetown University, Washington, DC. Unpublished manuscript.

Johnson, William and Derek Neal. 1998. "Basic Skills and the Black-White Earnings Gap." Pp. 480–500 in *Black-White Test Score Differences,* edited by C. Jencks and M. Phillips. Washington, DC: Brookings Institution.

Kirschenman, Joleen and Katherine Neckerman. 1991. "'We'd Love to Hire Them, but. . .': The Meaning of Race for Employers." Pp. 203–234 in *The Urban Underclass,* edited by C. Jencks and P. Peterson. Washington, DC: Brookings Institute.

Moss, Philip and Chris Tilly. 2001. *Stories Employers Tell: Race, Skill, and Hiring in America.* New York: Russell Sage.

Neal, Derek and William Johnson. 1996. "The Role of Premarket Factors in Black-White Wage Differences." *Journal of Political Economy* 104:869–95.

Oettinger, Gerald S. 1996. "Statistical Discrimination and the Early Career Evolution of the Black-White Wage Gap." *Journal of Labor Economics* 14:52–78.

O'Neill, June. 1990. "The Role of Human Capital in Earnings Differences between White and Black Men." *The Journal of Economic Perspectives* 4(4):25–45.

Pager, Devah. 2003. "The Mark of a Criminal Record." *American Journal of Sociology* 108:937–75.

———. 2007a. "The Use of Field Experiments for Studies of Employment Discrimination: Contributions, Critiques, and Directions for the Future." *Annals of the American Academy of Political and Social Sciences* 609:104–133.

———. 2007b. *Marked: Race, Crime, and Finding Work in an Era of Mass Incarceration.* Chicago, IL: University of Chicago Press.

Pager, Devah and Diana Karafin. 2009. "Bayesian Bigot? Statistical Discrimination, Stereotypes, and Employer Decision-Making." *Annals of the American Academy of Political and Social Science* 621(1):70–93.

Pager, Devah and Lincoln Quillian. 2005. "Walking the Talk: What Employers Say Versus What They Do." *American Sociological Review* 70(3):355–80.

Petersen, Trond, Ishak Saporta, and Marc-David L. Seidel. 2000. "Offering a Job: Meritocracy and Social Networks." *American Journal of Sociology* 106:763–816.

Pettit, Becky and Bruce Western. 2004. "Mass Imprisonment and the Life Course: Race and Class Inequality in U.S. Incarceration." *American Sociological Review* 59:151–69.

Roscigno, Vincent J., Lisette Garcia, Sherry Mong, and Reginald Byron. 2007. "Racial Discrimination at Work: Its Occurrence, Dimensions, and Consequences." *The New Black: Alternative Paradigms and Strategies for the 21st Century Research in Race and Ethnic Relations* 14:131–55.

Steele, Claude M. and Joshua Aronson. 1995. "Stereotype Threat and the Intellectual Test

From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Performance of African Americans." *Journal of Personality and Social Psychology* 69(5):797–811.

Tomaskovic-Devey, Donald, Melvin Thomas, and Kecia Johnson. 2005. "Race and the Accumulation of Human Capital across the Career: A Theoretical Model and Fixed-Effects Application." *American Journal of Sociology* 111(1):58–89.

Trope, Yaacov and Erik P. Thomson. 1997. "Looking for Truth in All the Wrong Places? Asymmetric Search of Individuating Information about Stereotyped Group Members." *Journal of Personality and Social Psychology* 73(2):229–41.

Turner, Margery, Michael Fix, and Raymond Struyk. 1991. *Opportunities Denied, Opportunities Diminished: Racial Discrimination in Hiring.* Washington, DC: Urban Institute Press.

Waldinger, Roger and Michael I. Lichter. 2003. *How the Other Half Works: Immigration and the Social Organization of Labor.* Berkeley, CA: University of California Press.

Western, Bruce and Becky Pettit. 2005. "Black-White Wage Inequality, Employment Rates, and Incarceration." *American Journal of Sociology* 111:553–78.

Whatley, Warren C. 1990. "Getting a Foot in the Door: 'Learning,' State Dependence, and the Racial Integration of Firms." *The Journal of Economic History* 50(1):43–66.

Wilson, William Julius. 1996. *When Work Disappears: The World of the New Urban Poor.* New York: Vintage Books.

Yarkin, K. L., J. P. Town, and B. S. Wallston. 1982. "Blacks and Women Must Try Harder: Stimulus Person's Race and Sex Attributions of Causality." *Personality and Social Psychology Bulletin* 8:21–24.

From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Bayesian Bigot? Statistical Discrimination, Stereotypes, and Employer Decision Making

By
DEVAH PAGER
and
DIANA KARAFIN

Much of the debate over the underlying causes of discrimination centers on the rationality of employer decision making. Economic models of statistical discrimination emphasize the cognitive utility of group estimates as a means of dealing with the problems of uncertainty. Sociological and social-psychological models, by contrast, question the accuracy of group-level attributions. Although mean differences may exist between groups on productivity-related characteristics, these differences are often inflated in their application, leading to much larger differences in individual evaluations than would be warranted by actual group-level trait distributions. In this study, the authors examine the nature of employer attitudes about black and white workers and the extent to which these views are calibrated against their direct experiences with workers from each group. They use data from fifty-five in-depth interviews with hiring managers to explore employers' group-level attributions and their direct observations to develop a model of attitude formation and employer learning.

Keywords: racial discrimination; employment; employer interviews; African Americans; stereotypes

The continuing significance of race in the minds of employers has been demonstrated in numerous contexts. Interviews with employers reveal the persistence of strong negative associations with minority workers, with particularly negative characteristics attributed to African American men (Kirschenman and Neckerman 1991; Moss and Tilly 2001; Wilson 1996). Studies of hiring behavior likewise suggest that employers strongly prefer white (and Latino) workers to otherwise similar African

NOTE: We thank Bruce Western for his contributions to the original interview schedule and for valuable comments on the article. Thanks also to Susan Fiske and Jeremy Freese for helpful comments and suggestions. Support for this project came from grants from NSF (SES-0547810), NIH (K01-HD053694), and a William T. Grant Foundation Scholar's Award. Direct all correspondence to Devah Pager, Department of Sociology, Princeton University, Princeton, New Jersey 08544; e-mail: pager@princeton.edu.

DOI: 10.1177/0002716208324628

NIJ 2005-IJ-CX-0019

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on January 15, 2009

From callbacks to labor markets - (in-person audit, New York City (Pager and Western 2009) 765

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Americans (Pager 2003; Pager, Western, and Bonikowski 2007; Bertrand and Mullainathan 2004; Bendick, Brown, and Wall 1999; Fix and Struyk 1993). Where the continuing existence of discrimination is a matter of little controversy, however, the underlying causes remain widely contested.

Much of the debate over the causes of discrimination centers on the rationality of employer decision making. Economic models of statistical discrimination, for example, emphasize the cognitive utility of group estimates as a means of dealing with the problems of uncertainty (Phelps 1972; Arrow 1972). Group-level estimates of difficult-to-observe characteristics—such as productivity, reliability, or willingness to submit to authority—can provide useful information in the screening of individual applicants. If employers can accurately estimate differences in the skills or disposition of blacks and whites on average, this information can be helpful in guiding decisions about individual black and white candidates for whom these characteristics are more difficult to observe directly.

Sociological and social-psychological models, by contrast, question the degree to which group-level attributions reflect accurate assessments (Bielby and Baron 1986; Tomaskovic-Devey and Skaggs 1999). Although mean differences may exist between groups on some productivity-related characteristics, these differences may be inflated in their application, leading to much larger differences in individual evaluations than would be warranted by actual group-level trait distributions (Rothschild and Stiglitz 1982).[1] Furthermore, estimates of group characteristics may reflect outdated associations, with factors such as occupational segregation, imperfect information flows, and negative feedback effects reducing awareness of changing distributions (Whatley and Wright 1994; Arrow 1998; Farmer and Terrell 1996). It thus remains unclear whether employers' assessments of various racial groups represent accurate representations and to what extent these assessments are responsive to novel or competing sources of information.

In this study, we take one step toward investigating these questions by measuring the nature of employer attitudes about black and white workers and the extent to which these views are calibrated against their direct experiences with workers from each group. Using data from fifty-five in-depth interviews with hiring managers, we explore employers' group-level attributions and their direct observations to develop a model of attitude formation and employer learning.

Devah Pager is an associate professor of sociology and a faculty associate of the Office of Population Research at Princeton University. Her research focuses on institutions affecting racial stratification, including education, labor markets, and the criminal justice system. Her current research has involved a series of field experiments studying discrimination against minorities and ex-offenders in the low-wage labor market. Recent publications include Marked: Race, Crime, and Finding Work in an Era of Mass Incarceration *(University of Chicago Press 2007) and "Walking the Talk: What Employers Say versus What They Do" (with Lincoln Quillian), in* American Sociological Review *70, no. 3 (2005): 355-80.*

Diana Karafin is a PhD candidate at Ohio State University whose research focuses on consequences of neighborhood integration, racial democracy and crime, and discrimination in the housing and labor markets.

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on January 15, 2009

From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Prior Research on Discrimination

Direct observations of employer behavior suggest that race continues to shape employment opportunities in important ways. A recent field experiment of employment discrimination in New York City, for example, provided a rare glimpse into the pervasiveness of discrimination in low-wage labor markets (Pager, Western, and Bonikowski 2007). Across hundreds of applications for entry-level jobs, blacks were half as likely to receive a callback or job offer as equally qualified white applicants. Furthermore, blacks with clean records fared no better than a white man just released from prison. The results of this and earlier audit studies provide vivid illustration of the degree to which racial considerations continue to actively shape the employment opportunities available to young black men (Bendick, Brown, and Wall 1999; Fix and Struyk 1993; Bertrand and Mullainathan 2004; Pager 2003; see Pager [2007] for a review).

Although the foregoing studies provide an important measure of the demand side of the labor market, audit methods offer little insight into the motivations or attitudes that shape employer behavior. Does discrimination typically reflect racial animus? Have employers had negative experiences with African American employees in the past that have led them to shy away from hiring blacks? Or do other factors shape employer decision making?

Previous research leaves these questions only partially answered. The classic study by Kirschenman and Neckerman (1991) indicates that race is indeed salient to employers in their hiring decisions. Results from a large sample of in-depth interviews reveal employers' blatant admission of their avoidance of young inner-city black men, attributing characteristics such as "lazy" and "unreliable" to this group (p. 213; see also Wilson 1996; Moss and Tilly 2001; Waldinger and Lichter 2003).[2]

At the same time, while this study has been widely cited as evidence of employers' deep biases about African Americans, especially young black inner-city men, the findings themselves present a more complicated picture. While some employers spoke only in general terms about the assumed characteristics of black inner-city men, suggestive of the role of broad cultural stereotypes, others made specific reference to negative experiences with their own black employees, indicating that employer attitudes may be heavily shaped by direct observation of racial differences among their workers.

The underlying sources of employer attitudes thus remain somewhat unclear. Indeed, in framing their analysis with the concept of statistical discrimination, Kirschenman and Neckerman (1991) remain agnostic as to whether employers' comments represent accurate depictions or exaggerated stereotypes. Employers are clearly using race as a proxy for employment-relevant characteristics, but the degree to which the use of this proxy is informed by empirical realities remains uncertain.

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on January 15, 2009

From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.
767

Moss and Tilly (2001) also find employers readily referring to negative characteristics among African American workers, with pervasive concerns about dependability, motivation, attitude, and skill. Many of these employers cite concrete experiences with their own black employees as the basis of their attitudes, though some also cite media representations and more general observations of African Americans as the source of their racial attributions (pp. 138-40). At the same time, Moss and Tilly note that a "silent majority" of employers claim not to notice racial differences among their employees. The authors speculate that these responses are due to some combination of social desirability bias, effective screening or training techniques (rendering a population-level racial skills gap irrelevant for the particular firm), or an honest experience of black and white workers as comparable. Among the largest group of employers, then, it is difficult to assess to what extent perceptions of racial differences are relevant for employers' decision making.

While the present study can move no closer to assessing the accuracy (or honesty) of employers' racial assessments, it does attempt to provide a more explicit understanding of the connection between employers' direct experiences and their more general racial attitudes. Following in the footsteps of the important employer-interview studies from the 1990s, this project seeks to better understand the degree to which employers' extensive experience with black and white workers generates assessments of racial differences that reflect their own empirical observations. In exploring the link between experience and beliefs, we hope to provide some insight into the dynamic process by which racial attitudes are constructed and reinforced.

A Rational Actor Model of Hiring

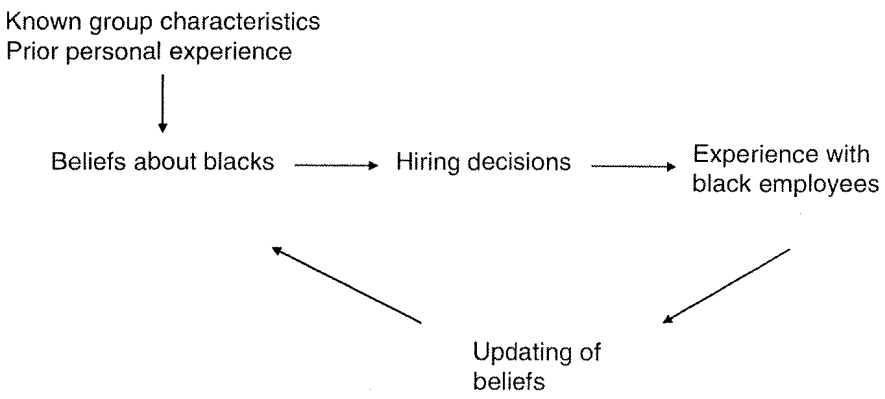
Economic models of statistical discrimination suggest that race offers a useful proxy for difficult-to-observe characteristics. Because productivity is difficult to observe directly, particularly for new hires, employers rely on indirect information inferred from group membership. This model can represent rational action on the part of employers, provided that the information they have about groups is accurate, and provided that there is a mechanism for updating estimates of group characteristics over time. Where discussions of statistical discrimination typically focus on accounting for single-point hiring decisions, the theoretical propositions can be readily extended to a dynamic process. Figure 1 represents such a model schematically, in its simplest form.

The model starts with information about *known group characteristics*: for example, employers may be aware of racial differences in graduation rates, test scores, incarceration rates, and other relevant disparities. Employers may also have direct *prior experiences* to which they refer in shaping their own beliefs. These antecedent sources of information then shape the employers' general *beliefs about blacks*—about their productivity, reliability, intelligence, and other

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on January 15, 2009

From callbacks to labor markets - (in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.
768

FIGURE 1
A RATIONAL ACTOR MODEL OF HIRING DECISIONS



relevant characteristics. Those beliefs will then guide individual *hiring decisions*. Once hiring decisions are made, employers are exposed to a range of black employees, and to the extent that these *experiences with black employees* do not fit the assumptions of the employer about members of that group, employers will revise their beliefs in a sort of Bayesian *updating process* to reflect a more accurate set of expectations.[3]

This equilibrium model of statistical discrimination describes the mechanisms by which employers can incorporate their direct observations into more general assessments of group characteristics. This feedback loop provides an efficient means of calibrating expectations, observations, and behavior. Consistent with such a model, some research does indeed point to evidence of employer updating. Altonji and Pierret (2001), for example, show that as firms acquire more information about a worker (through posthire observation), their evaluations (as reflected in wage offers) rely less on general (noisy) characteristics, such as educational attainment, and more on individual-specific characteristics, such as cognitive skill (see also Oettinger 1996; Farber and Gibbons 1996).[4] These studies provide compelling evidence that employers weigh their direct observations more heavily than inferences based on group proxies and that employer learning can improve on initial estimates. Nonetheless, this line of research applies specifically to learning about individual workers, whereby group-based estimates are replaced with the observed characteristics of individuals. It remains unclear, however, whether an employer's learning about an individual employee affects the employer's expectations about the broader group to which that individual employee belongs.

Indeed, Farmer and Terrell (1996) provide an elegant theoretical analysis of employer learning and statistical discrimination in which initial employer beliefs are revised through an updating process similar to that described above.

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on January 15, 2009

From callbacks to labor markets - (in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.
769

According to the authors, however, the updating process might apply only to the specific employees under observation, rather than to members of the larger group. "Higher than expected output of one worker provides much information about individual ability, but only a single data point to estimate the average ability of a population of millions. In addition to observation of workers, an employer receives an abundance of information on average group ability from other sources. Observations of average output, or perhaps occupations, of other members of the group influence the assessment of group ability" (p. 206). The process by which employers generate estimates of group characteristics and update those estimates over time thus remains unknown. In the following analysis, we seek to make headway in understanding this dynamic process.

Source of Data

This research is based on in-depth, in-person interviews with 55 New York City employers. The employers in this study represent a subsample of firms advertising for entry-level positions in 2004.[5] In selecting respondents, we aimed to capture the full range of entry-level employers according to industry, occupation, and other types of characteristics thought to be associated with discrimination. In all, we made 243 in-person contacts with 152 firms to solicit participation in our study, 55 of whom agreed, rendering a response rate of 36 percent. The majority of respondents were male (70 percent); white (59 percent); located in Manhattan (82 percent); and managing a firm in the retail industry (46 percent), the restaurant industry (31 percent), or the service industry (11 percent).[6] Furthermore, 47 percent of respondents represented independent firms, 40 percent national chains, and 13 percent local New York City chains.

Interviews ranged in length from thirty minutes to two hours, with the average interview lasting fifty minutes. Content of the interviews reflected a wide range of topics, including recruitment strategies; screening procedures; concerns about entry-level workers; and criteria for selection, placement, and promotion decisions. The questions we focus on here come from a module focusing on employers' racial attitudes. This core segment of the interview probed three primary sets of issues related to (1) employers' general attitudes about the employment problems of black men, (2) their specific experiences with black applicants and employees, and (3) the relationship between employers' concrete experiences and their general attitudes. In the following discussion, we examine the pattern of responses that emerged from our conversations with employers, focusing specifically on the link between employers' general attitudes about blacks and their specific experiences.

The use of qualitative data is well suited to investigating complex processes given its ability to capture the nuance and depth of personal attitudes (Orbuch 1997). But this approach also has its limitations. The relatively small sample prevents us from drawing strong conclusions about the attitudes of employers more

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on January 15, 2009

From callbacks to labor markets - In-person audit, New York City (Pager and Western 2009) This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice. 770

generally. More important, the validity of findings from interview data depends on respondents' willingness to provide truthful answers to questions. Given the sensitive nature of this investigation, concerns over social desirability bias are highly relevant. As we discuss below, the candid answers we received from employers about their negative racial attitudes offer some reassurance that employers were not entirely self-censoring. At the same time, we must remain aware of the possibility that some responses may be affected by these concerns.

Despite these limitations, in-depth interviews provide a rare window into employers' thought processes and offer some leverage in understanding the complex process of attitude formation among this group. While we cannot conclusively adjudicate among theories based on the results from this study, we hope that our findings will be generative of hypotheses for further testing.

How Do Employers Think about Young Black Men?

In talking about race, and in particular the employment problems of black men, employers' responses represented a range of views.[7] Some employers emphasized the structural barriers facing African Americans, including poverty, a lack of education, disadvantaged neighborhood contexts, and prejudice and discrimination. For example, one employer from a courier company emphasized the problems of residential segregation, discrimination, and incarceration as key barriers to black men's employment opportunities.

> Well, there are of course the obvious problems of racial profiling where people don't want to hire them. They won't tell you that and you don't really think it happens as much in society. People think all of that segregation has ended, but it really hasn't. . . . Another problem, as a good example in terms of hiring people that have criminal backgrounds, a lot of people are not looking to take the chance, and unfortunately the number of African American men that have been incarcerated has just, in the past few years, been phenomenal. And so people don't stop to think that, let's say someone wasn't really involved in something. Maybe they just got caught up, so to speak. Some people may just think that, well you have been in trouble and so I don't want to hire you. So that is another aspect that keeps people from hiring black men.

Another employer from a retail chain pointed to a range of factors, from racism to a lack of education, that disadvantage black men in the labor market:

> Racism is still a huge issue in America. . . . Especially like black men have been repressed for so many years, like due to every, like due to racial issues, a lot of them are very like limited in their, their work experience, or in their education levels, because the education system in New York City is terrible, like in public schools especially, like in the inner city, and you know, in the poorer neighborhoods, so . . . there . . . is a lack of education . . . and so many people can't, you know, afford to go to college. . . . And therefore there's so many jobs that so many people are trying to compete for. . . . You know I think that that's what a lot of the problem is for, for, for especially black men in, in this city.

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on January 15, 2009

From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.
771

Overall, more than 40 percent of employers commented on structural issues, about 15 percent of whom emphasized these factors as their primary explanation for the employment problems of black men.

By far, the most common explanations for black men's employment problems, however, focused on the individual shortcomings of black men themselves. More than three-fourths of employers mentioned individual explanations at some point in their discussion of black men's employment problems, with well over half (60 percent) emphasizing individual factors as their primary explanation. A wide range of perceived shortcomings were identified as primary explanations for black men's employment troubles, including concerns over work ethic, attire, and attitude, which we discuss in detail below.

Of course, we know from other research that Americans tend to give individualistic explanations for inequality in general, whether racial or otherwise (Schuman and Krysan 1999; Kluegel and Smith 1985; Jackman 1994; Bobo 2004). For example, in his analysis of General Social Survey data collected between 2000 and 2004, Hunt (2007) finds that at least half of Americans believed contemporary racial inequality to be caused by a "lack of motivation," relative to less than a third who cited discrimination as an important problem.[8] To some extent, then, the distribution of responses we observe reflects a more general reluctance to view inequality in structural terms.

At the same time, however, the content in these interviews goes beyond the standard narrative of "pull yourself up by your bootstraps." Certainly we also heard comments of this sort. One employer from a local restaurant chain, for example, insisted, "If you are persistent, something will eventually pan out for you. I am sure of it. If you really want a job, you will eventually find one." The comments we focus on in this analysis, by contrast, go well beyond these generic beliefs in individualism to reveal far more specific attributions about black men.

In asking employers to reflect on their experiences with workers from different racial groups, we are not simply asking them about their general beliefs about inequality; we are asking them to draw from their expertise as *employers* to help us better understand why the economic outcomes for some groups are systematically better/worse than others. Given their unique vantage point, we might expect employers to express attitudes about the characteristics of black and white workers that diverge from mainstream American racial attitudes. Because of highly segregated social networks, many white Americans' exposure to African Americans is limited to casual observation, brief encounters, and media representations. Most employers, by contrast, have had extensive contact with black workers and have had the opportunity to observe these workers perform specific tasks and responsibilities.

Under these conditions, then, we might expect employers' direct experiences to play a larger role in shaping their racial attitudes relative to other white Americans. Previous research provides only partial insight into this question. Bobo, Johnson, and Suh (2002), for example, find that employers' racial stereotypes are indistinguishable from those of the general public, suggesting that workplace power or experience does little to shift generalized racial associations.[9]

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on January 15, 2009

From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.
772

At the same time, close-ended survey measures capture only one dimension of attitudes, potentially overlooking some of the complexities and contradictions embedded in employers' ideas about the characteristics of racial groups. The present study uses in-depth interviews to investigate employers' attitudes about and experiences with black men. We begin with an investigation of employers' general characterizations of black men, and then explore the degree to which these characterizations are rooted in direct experiences and observations.

Lack of a work ethic

One of the most common themes we heard from employers centered on the perceived lack of a work ethic among black men (fully 55 percent mentioned this issue). Some of these employers referred to a general lack of motivation to work among African American men. Others described a desire among black men to take advantage of the system instead of working. One employer at a retail store said simply, "I will tell you the truth. African Americans don't want to work." A manager of a retail store said similarly, "They don't want to work—you can tell by the attitude, clothing, the general body language." The owner of a dry-cleaning store commented, "They just don't have any drive. No get up and go attitude." Likewise, a young male employer working in a national clothing chain stated, "I think for a lot of them they are too lazy to work so they are not going to work."

A human resources manager at a national retail chain, when asked what she thinks explains the employment problems of black men, explained,

> Employer (E): They are not as motivated as other races to, you know, to get out and do more and to improve or make themselves better. ... [T]hey are not as motivated or determined to move up or to even do anything to improve themselves.
> Interviewer (I): What do you think this lack of motivation stems from?
> E: I don't know how to say this but probably laziness. Just being lazy and not wanting to work.

When asked the same question, the manager of a local restaurant chain in New York City first argued that more black men are employed than the statistics indicate because many work off the books. However, the employer then refined his initial explanation, stating, "Listen, I also think there is a degree of being lazy." As with many of the other employers interviewed, this manager seemed to believe that if black men were motivated to search for employment in a responsible manner, their employment problems would be resolved.

In several cases, employers attributed the lack of work ethic to a complex history of racism and paternalism. For example, the long-term manager of an industrial supply store in the Bronx noted, "In America blacks believe the golden opportunity is to be taken care of." Another employer who earlier had asserted that "African Americans don't want to work" elaborated on his initial comment:

> Maybe they think that this country owes them so much. Because of slavery and all of that. They, they tend to forget that was a couple of years ago. Instead of catching up with

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on January 15, 2009

From callbacks to labor markets - (in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.
773

> the world, they still keep that anger. . . . So African American men feel like they deserve something. That is basically why they don't want to work. They feel that if they can get things for free, why not?

The manager of a national restaurant chain expressed a similar view, differentiating between black men and black women. "I think the mentality is changing. I think [black men] are more accepting of letting the female work and they stay home." He went on to explain,

> I think, you know, a percentage of it, probably has to do with them figuring out how to take advantage of the system. You know, with welfare or something like that. I do think that a lot of them have the mentality of: I'm black and the government doesn't help. So, let me do what I can to get from the government. Though Martin Luther King has changed the world, a lot of them have the mentality that it is still a white world.

A different employer emphasized that programs exist to help those in need of work but that black men elect to ignore the programs as they want "things handed to them." The main supervisor of a national food chain explained,

> The key is, I think there are great programs out there, but people don't take advantage of them. It is probably laziness. You know, there are programs. But people are lazy. A lot of people want things handed to them.

Another employer focused more specifically on negative consequences of welfare dependence:

> Unfortunately we've bred generations of welfare—of people whose . . . careers are . . . welfare and social agencies. And, uh, and it's unfortunate. A lot of these people just don't have any work values.

Applying a similar line of reasoning, the young manager of a local telemarketing firm expressed frustration with black men playing the "race card" as a way of escaping responsibility:

> I mean, I do understand that sometimes the black man is racially profiled. So it is something they are always going to deal with. This is unfortunate for them, but I think they shouldn't even go there with playing the race card. I mean, a lot of them are like, "You treat me this way because I am black," and "I can't do this and I can't do that." I just think for a lot of them, it is that they just don't want to do things. They want to smoke weed and be a rapper.

Although this employer does see racial discrimination as a reality, he views it as more often used as a convenient excuse for failure than a direct cause. Underlying the employment problems of black men, according to this employer, is more often a simple lack of motivation and effort.

According to each of these employers, a lack of work ethic, motivation, and personal responsibility is pervasive among African American men, and together,

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on January 15, 2009

From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009) This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice. 774

these factors represent the primary causes of this group's employment problems. Historical social policies (e.g., slavery and welfare) are seen less as explanations for these patterns and more as convenient excuses available for African Americans in the abdication of responsibility.

Self-presentation

A second persistent theme in the interviews was the issue of black men's self-presentation (45 percent of employers mentioned this issue). In particular, employers highlighted problems of unsuitable appearance, negative attitude, and inappropriate conduct among black men. Particularly for positions involving customer service, employers expressed concern about the image projected by black men in their attire and attitude. An employer representing a local New York City clothing chain remarked,

> Sometimes these people looking for a job, why would they have a do-rag on and jeans? Why would they? You know, just the way you present yourself, it's how you are. . . . The black male, yes. It's just the way they carry themselves.

Another employer, the manager of a children's clothing store, when explaining why black men have difficulty securing employment, stated,

> The way they present themselves in the store. If they come in, and excuse the word, but they are all thugged out, it is not somebody I want on my sales floor.

Also placing significant weight on the role of appearance in securing employment, an employer representing a popular local New York City retail chain commented,

> If you go out looking for a job with caps, baggy pants and triple x t-shirts or whatever, you can't expect someone to hire you like that. Why don't you put on some decent clothes and go look for a job? That is the mistake minorities, black and Hispanics tend to do. They look for a job, and when they don't get hired, they automatically say it is discrimination.

The manager of a retail clothing chain complained, "[Black men are] usually, like very urban looking, baggy pants, you know, just like baggy clothes, hat." Likewise, an employer for a moving company said, "I think people who come in wearing baggy pants or something like that just make a bad impression. You can be green, orange, purple, or whatever. It doesn't matter, it isn't good."[10] For these employers, the problems of attire—and specifically the issue of baggy pants—signaled a lack of professionalism or an ignorance about appropriate work attire, something they often associated with black men.

In addition to matters of appearance, employers' comments about the self-presentation of black men also emphasized issues of attitude and conduct. Black men were often perceived as having a "bad attitude" about work or relations of

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on January 15, 2009

From callbacks to labor markets - (in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.
775

authority, or at times behaved in ways considered inappropriate for mainstream work environments. When considering the employment problems of black men, an employer from a national clothing chain explained, "It's just the way they carry themselves." Another described black men's tendency to present "language and the attitude, like gangsta or street."

The manager of a large national retail chain noted that differences in the self-presentation and conduct of black and white men affected the way they are perceived by employers and customers:

> I have heard people say that it is easier to deal with a white person more than a black person. I guess because they feel black people are always loud and hyper. Which is true. White people may know how to carry themselves a little better than black people. Black people always want to make a scene and always want to be heard.

A manager of a national retail store emphasized the distinction between qualifications and self-presentation, with the latter undermining the former in the case of black men:

> You know, I know a lot of black males and I know how they react to things and I know why they don't get jobs. I kind of know why. Some of them are actually qualified for jobs but they go in with that attitude. It is all in how they present themselves.

Several employers commented on the attitude problems they perceived among young black men as reflecting a sense of hostility or resistance to authority. An employer from a small, independent retail store stated, "[Black men have] this kind of attitude that is, is umm, resentful. It is hard. It is not an attitude that is favorable for business." An employer representing a large national clothing chain emphasized that black men don't present themselves to employers appropriately because they have a chip on their shoulder:

> I: Why do you think they are not presenting themselves in what you think is the best way? Where does this stem from?
> E: Um, ego and insecurity. The insecurity part is that they don't feel that they are right for the job. They lack that confidence but yet their ego won't let them admit it so they have a chip on their shoulder.

The owner of a local garment factory echoed some of the same concerns:

> [Black men] act a little more belligerent than others. There is the attitude and pattern of animosity with this group. They have a chip on their shoulders. They think, man, you are white and that is why you don't give me a job. That kind of thing. . . . The black male will come and say I am better than this and better than that and so on. So there is more of a macho type of attitude with the black male. When you go for a job you have to be, besides appearing decent and trustworthy, you can't look macho or act like you are better than other people.

Another employer at a retail store found it difficult to advise young black men about appropriate attire because of their resistance to authority.

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on January 15, 2009

From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.
776

> I think that it is hard for these men because they are too proud. Especially the younger guys, the eighteen- and nineteen-year-old black guys that come in here, are like, "Who are you to tell me no?" They have a real problem with authority.

These employers perceive black men to exhibit styles of dress or demeanor that present barriers to their mainstream involvement in the labor market. Concerns about the self-presentation of black men, ranging from baggy pants to bad attitudes, were viewed as a key obstacle to employment for this group.

A threatening or criminal demeanor

A final major theme in employers' general comments about black men focused on a perceived threatening or criminal demeanor. Extending comments about the "hardness" or "animosity" of black men discussed above, roughly a third of employers mentioned concerns of this sort. According to an employer from a local garment factory,

> I find that the great majority of this minority group that you are talking about either doesn't qualify for certain jobs because they look a little bit more, they come on as if, well, they are threatening.

An employer from a popular national clothing chain dismisses the racial frame initially but ultimately reached a similar conclusion:

> I don't know if I consider it on a race level. I just consider it more on approachability. And if someone seems intimidating, you know, and which, stereotypically might be, you might consider like a Hispanic person or a black person more intimidating than like a white person.[11]

Other employers focused more specifically on concerns about criminality. According to one employer from an ice cream chain, for example,

> I notice here working in the store, sometimes, a group of young black males will come in. And sort of, a red flag goes up. Everything stops, and you wonder, what are they going to do? Are they here to buy something? There is a general belief that, because of the way they dress and how they carry themselves, that they are trouble.

An employer from a retail store explained, "I mean, black males are not expected to go out and work, because they are doing other illegal things in the neighborhood." Likewise, an employer for a moving company pointed to the lure of criminal activity among African Americans:

> They see the drug dealer who is driving around in a Lexus, and then they see me in a Chevy, and they say I don't need to be a mover. I don't need to work sixty hours a week. I can go hustle stuff on the corner and drive a Lexus.

Another explained simply, "Half of them are in jail." These employers associate black men with danger, crime, and the criminal justice system, factors that appear incompatible with legitimate work.

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on January 15, 2009

From callbacks to labor markets - In-person audit, New York City (Pager and Western 2009) This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice. 777

Overall, then, we find fairly pervasive negative attitudes about black men as a group. While some employers did place more emphasis on lack of opportunity, prejudice, or other more structural factors, the majority of employers pointed to specific deficiencies among black men that led them to have difficulty in the job market. Our next set of questions seeks to investigate to what extent these negative attitudes are a function of employers' direct experiences.

Are Employers' Perceptions Informed by Experience?

In discussing their general attitudes about black men, many employers inevitably referred to specific observations or experiences in their own workplace. These anecdotes point to one potential source of employers' general attitudes, though they tell us little about the overall distribution of experiences. Our next set of interview questions, then, attempted to elicit more systematic information about employers' experiences with applicants and employees from various racial groups.[12] In particular, we asked employers to consider the extent to which the characteristics they had described to us were reflected in the black applicants or employees they had encountered over the past year.

In some cases, there was clearly a link between employers' general attitudes and their direct experiences. For example, one employer who had expressed negative views of blacks generally ("basically these people are lazy") went on to describe the differences he has observed between black and white applicants.

> Black people, mostly, yeah I can say that they come in, and "Are you hiring?" That's their question, and then that's it. They don't have, they are not prepared at all. I give them an application and they don't show up back with it. It's like, "I'm taking it, but let's see what happens. . . ." White people . . . they mostly come with resume already done, the paper with them.

Another employer who had commented earlier on dysfunctional culture affecting "people that come from two or three generations that are on welfare" described the differences he observed:

> Well for my business I am looking for somebody that comes dressed for the interview. If you walk in with baggy jeans it is not even worth talking to you. This is something I have observed with black men. I mean, yeah, I definitely have noticed that a lot of black male applicants typically don't know how to properly talk to me, or they leave messages on the phone that aren't really what I am looking for. I mean, they just don't seem energetic or like they really want this job.

An employer from a national service organization commented on her experiences with black applicants who seemed to take job opportunities for granted, not demonstrating suitable interest or seriousness of purpose.

> People will call here and you will try to schedule them for an interview. So you will offer them an interview, and they call up asking for one, and they say things like, "Yeah, I don't know if that time works for me." You know what I mean? It is just not proper etiquette.

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on January 15, 2009

From callbacks to labor markets - (in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.
778

While we were conducting an interview with the manager of a local telemarketing firm, a black male job applicant entered the room, dressed in baggy clothing, and asked to schedule an interview. The manager replied that the individual would have to come back later in the week, to which the applicant stated, "Gotcha. See you then." After the applicant left, the manager stated, "Not gonna hire him. See what I mean? Where did he learn that 'gotcha' was the right thing to say to a potential boss? And see what I mean about attire?"

A number of the employers we spoke to reported at least some negative experiences with black applicants and/or employees, ranging from comments about individual workers to descriptions of more general patterns. Overall, among our sample of white respondents, employers were more likely to note racial difference among their applicant pool (46 percent) than among their employees (34 percent).[13] This suggests that employers are doing an effective job at weeding out good employees from a more heterogeneous applicant pool. At the same time, note that more than half of employers claim not to notice racial differences in the quality of their applicants, and fully two-thirds of employers notice no racial differences among their own employees. In these cases, there is often little relationship between the impressions they report about African Americans generally and their own direct observations of black employees.

For example, one employer earlier emphasized the lack of work ethic among blacks ("just being lazy and not wanting to work") as the major factor for their employment problems; but when asked if she had observed these problems in her workplace, she replied, "Well no. . . . Of course once in a while they goof off, but that is across the board. I don't see any differences between groups." A video store employer acknowledged that stereotypes were often quite different from reality: "There are people that say there are differences in work ethics of black men, but I have no specific thing like that that I have noticed. Absolutely not. The worst employees have been the non people of color. They have been the worst."

An employer who earlier had alluded to the problems black men have with "presentation" and "attitude," went on to describe his employees with the following comments:

I: Among your employees, have you noticed differences in the work performance of blacks and whites?
E: Um, do you mean in terms of work ethic?
I: Sure.
E: They really have a nice work ethic.
I: What about comparing whites to blacks?
E: In my experience blacks will outdo them.
I: Blacks will outdo whites?
E: Yeah, once you get the right person. Sometimes with an entry-level, they don't seem as committed.
I: You mean your white employees?
E: Yeah. It is like they think they are above this. I don't find this with the black employees.

An employer for a retail clothing store (who had referred to blacks as being more "intimidating" than whites) came to a similar assessment about her own employees: "It's hard to compare because it's different types of work that they're doing,

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on January 15, 2009

From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.
779

but I would say that the people who are not of Caucasian descent work a little harder than the white kids."

In several cases, employers did acknowledge differences among workers but accounted for them along nonracial lines. We asked an employer who had earlier emphasized the importance of attire for her workers:

I: Do you observe that black male applicants are more likely than white male applicants to present themselves to you inappropriately?
E: No not really. I think it is a problem for all of them. I think just men in general.

In another case, the employer started out describing the problematic applicants she deals with. When we pressed her to consider whether these characteristics were more prevalent among black men, she thought about the question and then reframed her comments in terms of age.

E: They don't come in dressed, they don't come in and speak to you in a correct way. They speak like they are hanging out and not looking for a job.
I: Okay, so do you notice that black males come in not prepared or dressed appropriately more often than white males that are looking for jobs?
E: I think it is about the same honestly. It is more about the age. The high school kids are the ones that don't come in dressed like they are looking for a job. As they get older you can see the difference because they are coming in in slacks and they are coming in in a shirt and they are speaking to you. They aren't just like, "Here is an application," and then they walk out the door.

These employers appear to have identified a cluster of behaviors or styles of presentation that signal poor performance, some of which are often associated with African American men, but when pressed to make sense of their observations, they focus on a different set of categories. Of course, it may be the case that employers feel uncomfortable making racial attributions and thus retreat to a language of gender or age out of social desirability concerns. We have no way of conclusively ruling out this possibility. At the same time, given the extremely candid racial remarks made by these employers just moments before, we feel some reassurance in taking these responses at face value. These comments lead us to believe that employers felt sufficiently at ease to speak in racial terms, and thus, we see little reason that they would suddenly retreat into a more politically correct style of discourse. In fact, we were concerned with the opposite effect, that employers would feel some pressure to come up with specific examples of poor performance among their black employees, if only to maintain consistency with their earlier comments. Quite the contrary, we found employers repeatedly emphasizing the lack of clear distinctions between black and white applicants and employees, even as these comments appear to contradict earlier statements about the generalized characteristics of black men.

As a final example of the disconnect that often appeared between employers' expectations and outcomes, one employer from a small retail store describes her recent experience with a black man she had hired for a stockperson position in her retail store:

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on January 15, 2009

From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009) This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice. 780

> Like, last year. I had this guy come in, with a big hat and a big jacket. I don't know if he had a criminal record, but he looked like it. But I was so desperate that I hired him. [Laughs] He was OK. But the way he dressed. Sometimes the way they dress. It is bad. I mean, a big hat. You try to be nice, but at the same time, I don't need a guy with a big hat. [Laughs] I was surprised that he worked out well. He finished the season very nicely."

In this case, despite the fact that the employer viewed this man as poorly dressed and potentially criminal, he ultimately wound up being a reliable employee. Indeed, employers frequently acknowledged that their first impressions of applicants were often completely off base. Here, the employer recognized her misjudgment, and yet, there is little sign that this experience caused her to rethink her more general, negative impressions of black men. Earlier in the interview, this employer had characterized black men as having attitude problems. "Socially there is a difference. In the neighborhood [black men] have a kind of attitude, that is, is um, resentful. It is hard. It is not an attitude that is favorable for business." Although here (and in other comments), she insists that her own experiences with black employees have been overwhelmingly positive, she retains strong negative impressions about black men "in the neighborhood," the source of which remains unknown.

[E]mployers frequently acknowledged that their first impressions of applicants were often completely off base.

How to Explain the Gap between Employers' Perceptions and Experience?

Whereas most of the employers in our sample expressed consistent negative attitudes about black men, far fewer could identify specific patterns or experiences among their own applicants or employees to fit these characterizations. How can we account for this surprising disconnect? While the results of this study cannot provide a definitive answer, we consider several plausible explanations.

The salience of negative events

First, it is possible that while employers' experiences with black applicants or employees *on average* may be similar to those with whites, a small number of negative experiences may hold especially strong weight in shaping attitudes

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on January 15, 2009

From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.
781

(Fiske 1998). Indeed, several of the employers in our sample referred to singular experiences that, while not representative of their experiences more generally, seemed to shape their associations of specific groups. According to one employer from a national retail chain,

> You know, everyone has a couple of bad hires. And you remember those very vividly. And who that person is can really impact. [He describes a negative experience with an African American female employee.] That person just stuck in my head. That was the first time I had done hiring during a holiday season, which is pretty stressful. And that person just stuck in my head. And I could see her. It was hard to not see her in other people that you meet.

According to this employer's account, one particularly negative experience with a black employee colored the employers' expectations of blacks in the future. Where this employer also noted a number of positive experiences he had had with African American employees, this negative experience appeared to carry especially strong weight.

In another case, an employer for a retail clothing store spoke about his negative experiences with a few black employees at a previous company.

> One of them in particular was threatening me and telling me after I fired him that he was going to wait for me outside and that he was going to get me. So that kind of thing sometimes gives you a vengeance. So, you know, you become biased a little bit.

Again, this employer acknowledges the ways in which a particularly unpleasant encounter contributed to his biases against black men more generally. It may be the case, then, that where employers may have had only a small number of unusually negative experiences with African Americans, these incidents may be the driving force behind generalized negative attributions. Benign or positive experiences create less salient memories, even if more prevalent overall.

At the same time, only a few of the employers in the sample reported extreme negative experiences, and more than half reported that their experiences with black and white applicants and employees were roughly comparable.[14] Without salient negative experiences coloring attitudes, what then might explain employers' persistent negative racial attitudes?

Selection and screening

A second possible explanation for the disconnect between employers' characterizations of black men in general and their direct experiences has to do with the various selection and screening procedures that weed out the most undesirable members of the group. Indeed, recall that employers were significantly more likely to report noticing racial differences in the characteristics of their job applicants than among their employees (46 versus 34 percent), suggesting that the hiring process leads to a more even distribution of productivity characteristics among black and white employees than exist in the general population. As one

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on January 15, 2009

From callbacks to labor markets - (in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.
782

employer mentioned, "Yeah, once you get the right person . . . ," indicating that the selection process can yield high-quality black employees, even if there is a great deal of variation within the general population. Recognizing that employees are not picked at random, we asked employers to comment separately on their perceptions of the applicant pool, expecting that racial differences may be more pronounced before the employers' active screening. The fact that more than half of employers claim not to notice racial differences even at this stage suggests that more remains to be explained than employer sorting.

Of course, the hiring process is not the only selection process at work. Showing up to apply for a job itself involves a process of selection, with the search process requiring some degree of organization, motivation, and a commitment to formal work. Particularly in recent years, as we have seen increasing numbers of young black men exit the labor force altogether (Holzer, Offner, and Sorensen 2005), selection into labor force participation may differentiate the types of black men employers encounter among their applicant pool from those in the general population.

This research does not make claims about employers' accuracy in detecting population-level characteristics relative to those observed in the workplace. Rather, we simply note that for many employers, whatever sources of information they use to infer general characteristics of black men, direct experiences with black applicants and employees do not appear dominant.

[F]or many employers, whatever sources of information they use to infer general characteristics of black men, direct experiences with black applicants and employees do not appear dominant.

Updating versus subtyping

A third potential explanation for the disconnect between employers' general attitudes and their specific experiences draws upon the social-psychological concept of subtyping. Where Bayesian models of updating assume that individuals incorporate new information by refining their expectations in ways consistent with their experiences, social psychological models emphasize the cognitive resistance to information that is disconfirming of expectations (Fiske 1998). A wealth of experimental evidence illustrates ways in which individuals are prone to view

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on January 15, 2009

From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009) This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice. 783

those who do not conform to stereotypes as exceptions, unrepresentative of the group as a whole, rather than as impetus for stereotype refinement (Taylor 1981; Weber and Crocker 1983). The creation of a subtype allows group stereotypes to remain intact in the face of disconfirming information.

Figure 2 applies the theory of subtyping to the schematic model proposed earlier to reveal a potentially different set of processes shaping racial attitudes. Here, instead of (or in addition to) known group characteristics and prior experiences being the predominant sources of racial attitudes, cultural stereotypes and media imagery play a strong role in shaping group expectations.[15] Those beliefs then translate into hiring decisions, which in turn provide a range of direct experiences for employers with black workers. But instead of positive experiences with black employees—as many of these employers report—leading to an updating of beliefs about blacks in general, we see a process of subtyping. As the saying might go, "*My blacks* aren't like blacks in general." Employers view their own black workers as a special subtype whose characteristics have little bearing on their evaluations of the larger group. Correspondingly, we see no pathway linking direct experiences with general beliefs.

Of course, this simple model does not capture the many nuances of attitude formation and change, which certainly contain a more diverse set of influences and pathways than are represented here. Rather, this basic schematic serves to represent one important alternative to the model proposed earlier, in which information flowed from the general to the specific and back again in ways that enhanced accuracy over time. Here, by contrast, an updating of expectations may occur with respect to specific members of the group (the subtype), but little revision of employers' more general attitudes is expected.

The analysis likewise does not imply that employers are necessarily wrong in their assessments of various group characteristics. It may be the case that employers' information about African Americans as a group is accurate and that the various selection processes at work (with individuals selecting into labor market participation and employers selecting workers) simply yield a more advantaged subgroup to which employers are exposed. This research makes no claims about the relationship between employers' attitudes and the "true" characteristics of African Americans. Rather, the findings suggest that, whatever sources of information they may be drawing from in forming racial attitudes, employers do not seem to draw heavily from their own personal experiences.

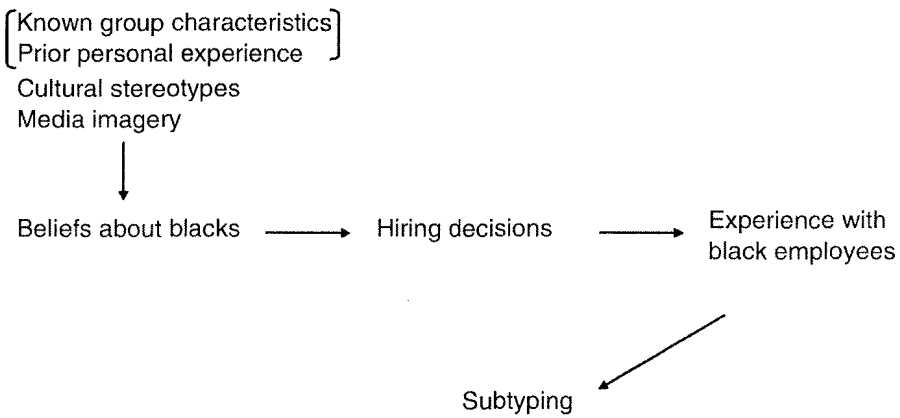
Conclusion and Discussion

The findings of this research suggest that, while most employers expressed strong negative views about the characteristics of African American men, fewer than half of these employers reported observations of their own applicants or employees consistent with these general perceptions. Where employers may

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on January 15, 2009

From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.
784

FIGURE 2
RESILIENCE OF STEREOTYPES



update their expectations regarding individual workers, these experiences do not seem to have noticeable effects on their attitudes about the group as a whole. Rather, employer attitudes appear more consistent with a model of subtyping, in which individuals who do not conform to a stereotype are viewed as exceptions whose characteristics have little bearing on the larger group.

Employers surely receive relevant information about various groups from sources other than direct workplace experience. They observe and interact with people in public spaces, they read newspaper coverage and watch TV news, and they are aware of racial inequality and have their own ideas about how and why this inequality is generated and maintained. These sources of information surely provide valuable complements to direct workplace experience. At the same time, it is surprising that the experiences employers report from their own direct observations do not carry greater weight in their general attitude formation. These results suggest that simple contact and exposure are themselves insufficient to revise deeply embedded racial attributions.

This analysis holds potentially troubling implications for hiring behavior. We know from the results of field experiments that employers consistently avoid black workers, hiring them at roughly half the rate of equally qualified whites. Where models of statistical discrimination might interpret this behavior as the rational response to observed differences in the productivity of black and white workers, the present research questions this conclusion. The majority of employers who report positive experiences with black workers (or no differences between black and white workers) nevertheless maintain strong negative attitudes about black men generally. To the extent that these attitudes shape hiring decisions, even in the scenario of equal productivity among black and white workers, we would expect the problems of hiring discrimination to persist well into the future.

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on January 15, 2009

From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.
785

Notes

1. See Armour (1997) for an extensive discussion of the logical, legal, and moral dilemmas of "reasonable racism."

2. Although the results of this study are indeed striking, it is important to keep in mind that more than 50 percent of Kirshenman and Neckerman's (1991) sample "either saw no difference [in the work ethic of whites, blacks, and Hispanics] or refused to categorize in a straightforward way" (p. 210).

3. A similar process of updating is described in the social psychological literature on stereotype change, referred to as a "book keeping model," according to which new information is incrementally incorporated into existing beliefs or attitudes about a group (Weber and Crocker 1983).

4. In this analysis, cognitive ability is observed to the researcher (by the respondent's score on an Armed Forces Qualification Test [AFQT] test measured earlier) but not by the employer at the point of hire. The assumption is that the worker's cognitive ability becomes observable to the employer with time on the job.

5. Roughly 80 percent of the employers in this sample were drawn from a random sample of employers advertising for an entry-level position in 2004 (see Pager, Western, and Bonikowski 2007). The remaining 20 percent were drawn through a purposive sampling technique enabling us to better represent large employers and industries underrepresented by the primary sample.

6. Note that within these industries are represented a wide range of job titles, including stockers, sales assistants, busboys, kitchen staff, waiters, couriers, and customer service positions.

7. The vast majority of employers offered multiple explanations within a single response series. In categorizing employers' sentiments, we distinguish between individual and structural explanations using several coding schemes. The first coding scheme takes into account the "first-mention," or whatever explanation was first proposed by the employer; the second takes into account any factor mentioned by an employer, with most employers being coded into several categories; the third uses a "holistic" approach in which we coded the comments according to what appeared to be the employers' main point. In many interviews, we specifically probed employers who reported multiple explanations with the question, "What do you think is the most important factor?" In other cases, this coding is based on our interpretation of the transcript. The main substantive conclusions are consistent across coding schemes, and where relevant, multiple measures are reported here.

8. An additional 10 percent cited "less inborn ability" as a major explanation, and 43 percent cited a "lack of chance for education." Respondents were allowed to choose more than one explanation, and thus, the categories do not sum to 100 percent (Hunt 2007, 400, Table 2).

9. Bobo, Johnson, and Suh's (2002) study draws data from the Los Angeles Study of Urban Inequality in which respondents were asked to rate members of a series of racial groups according to a range of characteristics (unintelligent, prefer welfare, hard to get along with, poor English).

10. Note that here (and in several other interviews), the job in question requires primarily manual work and few customer interactions. Jobs at a cleaning company or a moving company do not typically require professional dress, and yet, for these employers, a worker's attire sends an important signal about his reliability and commitment to work.

11. A number of other employers emphasized the perceptions of others about blacks as threatening or intimidating. For example, one employer said, "I think a lot of white people are scared of black people for some reason. I think they are scared of them, intimidated by them, they don't feel comfortable around them." Similarly, the manager of a small restaurant expressed concern over the aggressive demeanor of black men, stating that employers may be hesitant to call black applicants back because "immediately a black male is perceived as being aggressive." Although these perceptions may also be highly relevant for hiring decisions, we do not include these responses here as this analysis focuses on employers' own views about African Americans.

12. We asked separately about applicants and employees, given that effective employer screening would lead to a different distribution of characteristics among those seeking employment relative to those who become employed.

13. African American employers appeared slightly more inclined to notice racial differences among applicants but substantially less likely to report racial differences among employees. Immigrant employers noted more racial differences at both stages. Note, however, that our sample of African American and immigrant employers is small and that these indications must be taken as tentative.

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on January 15, 2009

From callbacks to labor markets - (in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.
786

14. Even respondents who described just one negative event were coded among those who had observed racial differences.

15. While we cannot measure the influence of cultural stereotypes directly, several employers did explicitly comment on the ways in which cultural representations of black men in the media affected their perceptions of black men. According to one employer, "I'd probably say 90 percent of the crimes you see on TV are African Americans, female or male, and that's something that's in the back of your head, you know." Similarly, another employer commented, "We have the media sending all this negative information about the young black male. All this negative information constantly. . . . We are getting the wrong image of what they look like."

References

Altonji, Joseph G., and Charles R. Pierret. 2001. Employer learning and statistical discrimination. *Quarterly Journal of Economics* 116:313-50.

Armour, Jody David. 1997. *Negrophobia and reasonable racism: The hidden costs of being black in America*. New York: New York University Press.

Arrow, Kenneth J. 1972. Models of job discrimination. In *Racial discrimination in economic life*, ed. Anthony H. Pascal, 83-102. Lexington, MA: D.C. Heath.

———. 1998. What has economics to say about racial discrimination? *Journal of Economic Perspectives* 12:91-100.

Bendick, Marc, Jr., Lauren Brown, and Kennington Wall. 1999. No foot in the door: An experimental study of employment discrimination. *Journal of Aging and Social Policy* 10:5-23.

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* 94:991-1013.

Bielby, William, and James N. Baron. 1986. Men and women at work: Sex segregation, and statistical discrimination. *American Journal of Sociology* 91:759-99.

Bobo, Lawrence D. 2004. Inequalities that endure? Racial ideology, American politics, and the peculiar role of the social sciences. In *The changing terrain of race and ethnicity*, ed. Maria Krysan and Amanda E. Lewis, 13-42. New York: Russell Sage Foundation.

Bobo, Lawrence D., Devon Johnson, and Susan A. Suh. 2002. Racial attitudes and power in the workplace: Do the haves differ from the have-nots? In *Prismatic metropolis: Inequality in Los Angeles*, ed. Lawrence D. Bobo, Melvin Oliver, James H. Johnson Jr., and Abel Valenzuela Jr., 491-523. New York: Russell Sage Foundation.

Farber, Henry S., and Robert Gibbons. 1996. Learning and wage dynamics. *Quarterly Journal of Economics* 111:1007-47.

Farmer, Amy, and Dek Terrell. 1996. Discrimination, Bayesian updating of employer beliefs and human capital accumulation. *Economic Inquiry* 34:204-19.

Fiske, Susan. 1998. Stereotyping, prejudice, and discrimination. In *The handbook of social psychology*, ed. Daniel Gilbert, Susan Fiske, and Gardner Lindzey, 357-411. New York: McGraw Hill.

Fix, Michael, and Raymond J. Struyk, eds. 1993. *Clear and convincing evidence: Measurement of discrimination in America*. Washington, DC: Urban Institute Press.

Holzer, Harry J., Paul Offner, and Elaine Sorensen. 2005. Declining employment among young black less-educated men: The role of incarceration and child support. *Journal of Policy Analysis and Management* 24:330-33.

Hunt, Matthew O. 2007. African American, Hispanic, and white beliefs about black/white inequality, 1977-2004. *American Sociological Review* 72:390-415.

Jackman, Man R. 1994. *The velvet glove: Paternalism and conflict in gender, class, and race relations*. Berkeley: University of California Press.

Kirschenman, Joleen, and Katherine Neckerman. 1991. We'd love to hire them, but . . . : The meaning of race for employers. In *The urban underclass*, ed. Christopher Jencks and Paul E. Peterson, 203-34. Washington, DC: Brookings Institution.

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on January 15, 2009

From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Kluegel, James R., and Eliot R. Smith. 1985. *Beliefs about inequality: Americans' views of what is and what ought to be*. New York: Aldine de Gruyter.

Moss, Philip I., and Christopher Tilly. 2001. *Stories employers tell: Race, skill, and hiring in America*. New York: Russell Sage Foundation.

Oettinger, Gerald S. 1996. Statistical discrimination and the early career evolution of the black-white wage gap. *Journal of Labor Economics* 14:52-78.

Orbuch, Terri L. 1997. People's accounts count: The sociology of accounts. *Annual Review of Sociology* 23:455-78.

Pager, Devah. 2003. The mark of a criminal record. *American Journal of Sociology* 108:937-75.

———. 2007. *Marked: Race, crime, and finding work in an era of mass incarceration*. Chicago: University of Chicago Press.

Pager, Devah, Bruce Western, and Bart Bonikowski. 2007. Discrimination in low wage labor markets. Working Paper, Office of Population Research, Princeton University, Princeton, NJ.

Phelps, Edmund. 1972. The statistical theory of racism and sexism. *American Economic Review* 62:659-61.

Rothschild, Micahel, and Joseph E. Stiglitz. 1982. A model of employment outcomes illustrating the effect of the structure of information on the level and distribution of income. *Economic Letters* 10:231-36.

Schuman, Howard, and Maria Krysan. 1999. A historical note on whites' beliefs about racial inequality. *American Sociological Review* 64:847-55.

Taylor, Shelley E. 1981. A categorization approach to stereotyping. In *Cognitive processes in stereotyping and intergroup behavior*, ed. D. L. Hamilton, 83-114. Hillsdale, NJ: Lawrence Erlbaum.

Tomaskovic-Devey, Donald, and Sheryl Skaggs. 1999. An establishment-level test of the statistical discrimination hypothesis. *Work and Occupations* 26:420-43.

Waldinger, Roger, and Michael Lichter. 2003. *How the other half works: Immigration and the social organization of labor*. Berkeley: University of California Press.

Weber, Renee, and Jennifer Crocker. 1983. Cognitive processes in the revision of stereotypic beliefs. *Journal of Personality and Social Psychology* 45 (5): 961-77.

Whatley, Warren, and Gavin Wright. 1994. Race, human capital, and labour markets in American history. In *Labour market evolution: The economic history of market integration, wage flexibility, and the employment relation*, ed. George Grantham and Mary MacKinnon, 270-91. London: Routledge.

Wilson, William Julius. 1996. *When work disappears: The world of the new urban poor*. New York: Vintage Books.

Downloaded from http://ann.sagepub.com at PRINCETON UNIV LIBRARY on January 15, 2009

From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009) 788

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Walking the Talk? What Employers Say Versus What They Do
Devah Pager; Lincoln Quillian
American Sociological Review; Jun 2005; 70, 3; ABI/INFORM Global
pg. 355

Walking the Talk? What Employers Say Versus What They Do

Devah Pager
Princeton University

Lincoln Quillian
Northwestern University
University of Wisconsin–Madison

This article considers the relationship between employers' attitudes toward hiring ex-offenders and their actual hiring behavior. Using data from an experimental audit study of entry-level jobs matched with a telephone survey of the same employers, the authors compare employers' willingness to hire black and white ex-offenders, as represented both by their self-reports and by their decisions in actual hiring situations. Employers who indicated a greater likelihood of hiring ex-offenders in the survey were no more likely to hire an ex-offender in practice. Furthermore, although the survey results indicated no difference in the likelihood of hiring black versus white ex-offenders, audit results show large differences by race. These comparisons suggest that employer surveys—even those using an experimental design to control for social desirability bias—may be insufficient for drawing conclusions about the actual level of hiring discrimination against stigmatized groups.

In 1930, Richard LaPiere, a Stanford professor, traveled twice across the country by car with a young Chinese student and his wife. The purpose of the trip, unbeknown to his travel companions, was to assess the reactions of hotel and restaurant proprietors to the presence of Chinese customers. During the course of 251 visits to hotels, auto camps, restaurants, and cafes, only once were they refused service. Six months later, LaPiere mailed a survey to each of the proprietors in which one of the questions asked, "Will you accept members of the Chinese race as guests in your establishment?" More than 90 percent of the respondents indicated unequivocal refusal. The discrepancy between these proprietors' responses to the surveys and their actual behavior is indeed striking: Although nearly none of the proprietors expressed a willingness to accept the patronage of Chinese customers, virtually all of them did so when confronted with the situation (LaPiere 1934). If we were to make generalizations based on either the survey results or the field study alone, we would develop radically different views on the level of racial hostility toward the Chinese at that time in history.

LaPiere's study provides a much needed reality check for researchers who rely on expressed attitudes for insight into the nature and causes of discriminatory behavior. Unfortunately, there have been very few efforts to provide the kind of comparison offered in LaPiere's study. Measures from surveys often are accepted as an adequate proxy for behaviors, with little effort to validate this assumption.

The current article seeks to make headway in this discussion, following up on the insights provided by LaPiere more than 70 years ago. In this discussion, we compare the self-reported

Direct all correspondence to Devah Pager, Department of Sociology, Princeton University, Princeton, NJ 08544 (pager@princeton.edu). Support for this research includes grants from the National Science Foundation (SES-0101236), the National Institute of Justice (2002-IJ-CX-0002), the Joyce Foundation, and the Soros Foundation. The views expressed in this document are those of the authors and do not necessarily represent those of the granting agencies. The authors are grateful to Harry Holzer, Michael Stoll, and Steven Raphael for sharing their survey instrument with us. The authors received helpful comments and suggestions from Bruce Western, Harry Holzer, Jeff Manza, Paul DiMaggio, Susan Fiske, and from the *ASR* editor and anonymous *ASR* reviewers.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

attitudes exhibited by a sample of Milwaukee employers with their actual behavior in real-life employment situations. By placing our analysis within the context of research on discrimination in contemporary labor markets we hope to underscore the degree to which method matters in our interrogation of the social world.

RACIAL ATTITUDES, DISCRIMINATION, AND CONTEMPORARY LABOR MARKET INEQUALITY

In the years since LaPiere's study, much has changed about race relations in the contemporary United States. In present times, it would be extremely rare to find respondents willing to state racial objections as candidly as those reported in LaPiere's survey. Indeed, trends in racial attitudes demonstrate steady movement toward the endorsement of equal treatment by race and the repudiation of direct discrimination. According to surveys conducted in the 1940s and 1950s, for example, fewer than half of whites believed that white students should go to school with black students or that black and white job applicants should have an equal chance at getting a job. In contrast, by the 1990s, more than 90 percent of white survey respondents endorsed the principle that white and black students and job applicants should be treated equally by schools and employers (Schuman et al. 2001).

Consistent with these trends, many indicators of social and economic status show that African-Americans have made great strides in approaching parity with whites. Blacks, for example, are now nearly equal to whites in rates of high school completion, and have become increasingly well-represented in occupational sectors previously dominated by whites (Farley 1997; Mare 1995; Wilson 1978). Likewise, in the decade after the Civil Rights Movement, and again during the 1990s, the wage gap between black and white workers was substantially reduced (Couch and Daly 2002; Harrison and Bennett 1995; but see Western and Pettit forthcoming). The rapid social mobility among blacks in the United States provides support for the notion that the progressive trends apparent in measures of racial attitudes reflect a real shift in the opportunities now available to African-Americans. In fact, these positive indicators have led some prominent academics to proclaim the problem of discrimination solved. Economist James Heckman, for example, has asserted that "most of the disparity in earnings between blacks and whites in the labor market of the 1990s is due to the differences in skills they bring to the market, and not to discrimination within the labor market." He went on to refer to labor market discrimination as "the problem of an earlier era" (Heckman 1998:101–102). Indeed, for many observers of contemporary race relations, the barrier of discrimination appears to have withered away, leaving blacks the opportunity to pursue unfettered upward mobility.

And yet, despite the many signs of progress, there remain important forms of social and economic inequality that continue to differentiate the experiences of black and white Americans. According to many indicators, blacks, and black men in particular, continue to lag far behind their white counterparts. Some indicators show black men doing steadily worse. African-Americans, for example, experience roughly double the rate of unemployment experienced by whites, with very little sign of change over time. Likewise, rates of joblessness among young black men have been rising over time (Holzer, Offner, and Sorensen 2005).

As a further troubling indicator, many of these young black men, instead of making their way through school and into jobs, are instead increasingly finding themselves housed in an expanding number of correctional facilities. Approximately 1 in 3 black men will spend some time in prison over his lifetime, as compared with only 1 in 17 white men (Bonczar 2003). Among young black high school dropouts, this figure rises to nearly 60 percent. Rivaling other conventional social institutions—such as military service, employment, and marriage, incarceration has now become a typical event in the life course of young disadvantaged men (Pettit and Western 2004).

How can we explain the discrepancies between these varied measures? On the one hand, the progressive trends in racial attitudes may reflect a genuine openness among white Americans to racial integration and equality. In this case, the continuing difficulties facing segments of the black population may simply reflect the "bumpy road" on an otherwise steady trajectory toward racial parity (Gans 1992).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.
790

Further, white racial attitudes are not the only barrier to black mobility. Changes in the economic structure, family composition, and crime policy, among other factors, may each exert an exogenous influence on the black population in ways that inhibit mobility, independent of the racial openness of contemporary institutions. From this perspective, continuing black disadvantage could be explained by a reasonable lag between changing attitudes and outcomes, as well as by the multiple influences that shape patterns of racial inequality.

On the other hand, traditional survey measures of racial attitudes may not accurately reflect the degree to which race continues to shape the opportunities available to African-Americans. Indeed, a great deal of evidence suggests that racial stereotypes remain firmly embedded in the American consciousness, affecting perceptions of and interactions with racial minorities even among respondents who overtly endorse the principle of equal treatment (Devine and Elliot 1995). Substantial levels of discrimination have likewise been detected by experimental field studies, which find consistent evidence of racial bias against black applicants in housing, credit, and employment markets (Bertrand and Mullainathan 2004; Turner, Fix, and Struyk 1991; Yinger 1995). As a further reflection of lived experience, the large majority of blacks continue to perceive discrimination as routine in matters of jobs, income, and housing (Feagin and Sikes 1994).

Given the available information, it is difficult to evaluate the extent to which direct discrimination plays a role in shaping the opportunities available to blacks in contemporary society. Surveys of racial attitudes portray one optimistic picture, whereas indicators of economic and social inequality present more mixed results. It is only through direct comparisons of these differing measures that we can assess how and why they may project such divergent conclusions.

In this article, we focus on the specific issue of employment discrimination. Substantively, we are interested in assessing the degree to which employer preferences or biases influence the opportunities available to stigmatized workers. Methodologically, we seek to assess the degree to which choice of measurement strategy affects our understanding of these processes. In our analysis of survey data and behavioral outcomes, we engage with LaPiere's central concern about the correspondence between measured attitudes and behaviors.

We begin with a review of the literature on the attitude–behavior relationship since LaPiere's study, focusing specifically on the case of attitudes toward and treatment of stigmatized groups. We then turn to concerns regarding the use of survey measures as proxies for measures of discrimination. Finally, we discuss the results from a matched field experiment and telephone survey that are the basis of our empirical results. Throughout this discussion, we seek to emphasize that investigations of important substantive concerns cannot be separated from a discussion of the methods by which these investigations take place.

ATTITUDES AND BEHAVIORS

Understandings of the Attitude–Behavior Relationship

Attitude questions are frequently asked because they are believed to be illuminating about one or more behaviors of interest. One of the most common uses of attitude research is to assess prejudices, stereotypes, and other measures of social distance with the goal of gaining insight about the nature of discriminatory behavior (National Research Council 2004). Attitude questions have been widely used as tools to assist in understanding the basis of behaviors such as discrimination in employment (Bobo, Johnson, and Suh 2000), residential mobility related to white flight (Farley et al. 1994, Krysan 2002), and the influence of race on voting patterns (Sniderman and Piazza 1993).

Because of the difficulty of gathering data on discrimination in natural settings, many substantive sociological studies of discrimination rely on easier-to-gather survey or interview data in the place of behavioral measures. Some studies focus on attitudinal indicators alone, leaving the connection to behavior implicit. Others ask respondents about past behavior or anticipated behavior in response to hypothetical situations. A wide range of survey scales and specific survey techniques have been developed to measure specific forms of prejudice and discrimination (National Research Council, 2004, chapter 8).

As one important example, a survey technique that has become increasingly popular for

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
From callbacks to labor markets - In-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

assessing situational discrimination involves use of the vignette question, which elicits reactions from respondents about fairly detailed hypothetical scenarios (Sniderman and Grob 1996). An influential early example of the vignette method was developed by Reynolds Farley and colleagues for a better understanding of the attitudinal sources of racial segregation (Farley, Bianchi, and Colosanto 1979; Farley et al. 1978; Farley et al. 1994). With Farley's approach, respondents are asked to express the level of discomfort they would experience living in hypothetical neighborhoods of varying racial compositions, and to estimate the likelihood that they would move out of such neighborhoods. Farley's innovative technique has become widely used in subsequent research, in part because it is easily combined with experimental survey techniques (discussed in the next section) (Emerson, Yancey, and Chai 2001; Krysan 2002).

A key assumption of vignette studies is that reported hypothetical behavior is an accurate proxy for the behavior that would be observed if the respondent actually encountered the situation. In the case of vignette studies that attempt to illuminate the process of white flight, for example, the assumption is that respondents who say they would feel discomfort or would move is highly related to the behavior of moving out (or not moving in) that would occur if the respondent actually lived in the hypothetical neighborhood. Although a perfect attitude–behavior correspondence is not required, these studies make the assumption that the two are related. An almost complete separation between attitudes and corresponding behaviors would undermine the rationale behind most attitudinal studies.

The expectation of attitude–behavior correspondence results naturally from the view that human action is the product of conscious mental states. Several psychologists, most notably Fishbein and Ajzen (1975), have formally modeled the relationship between these components to describe the formation of attitudes and their subsequent influence on behavior. In their model, feelings or beliefs about an object give rise to positive or negative evaluations of the object. These evaluations then influence behavioral intentions, which ultimately influence behavior (Ajzen 2001; Fishbein and Ajzen 1975). If attitudes can be measured successfully by survey questions or interviews, then these should have at least some power to predict overall patterns of behavior toward the attitude object.

Despite the clarity and intuitive appeal of this model, what is most striking about past investigations of the attitude–behavior relationship is the wide range of correlations reported across different studies. Both Deutscher (1966) and Wicker (1969), for example, review a number of studies that find virtually no relationship between attitudes and behaviors. Schuman and Johnson (1976) also discusses a number of notable studies in which a zero or negative correlation between attitudes and behaviors was found. In their review, however, they conclude that a majority of research on the attitude–behavior relationship finds a moderate positive relationship. With examples for each extreme, their article reports correlations close to zero among attitude–behavior assessments of racial bias and transient economic transactions, while demonstrating correlations exceeding .85 among studies of voting behavior. Most others are shown to fall somewhere in between (Schuman and Johnson 1976).

This literature supports the conclusion that no simple formula can describe the attitude–behavior relationship. Rather, tremendous variation exists in the measurement of attitudes and their associated behaviors, and assumptions about their correspondence should be reviewed with caution.

Attitudes And Behavior Toward Stigmatized Groups

Despite the appeal of using attitudinal measures as proxies for behavior, particularly for hard-to-measure behaviors such as discrimination, a number of factors complicate the relationship between verbal expressions on surveys and actual behaviors directed toward members of stigmatized groups. First, social surveys have long been plagued by the problem of social desirability, or the phenomenon that respondents seek to give socially appropriate answers to questions, even if this involves distorting the truth (Bradburn 1983). In the contemporary United States, the norms of racial equality are so strong that survey respondents are unlikely to feel comfortable expressing negative opinions about members of other racial groups (Crandall

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

1994). When asked questions about race or other sensitive issues, respondents may be led by these pressures to shift their opinions subtly (or in some cases not so subtly) in the direction of answers they perceive to be the most socially acceptable. To the extent that real-world discrimination continues, this has the effect of biasing survey results in the direction of politically correct, nonprejudicial responses, and of weakening the relationship between measured attitudes and behavior.

Researchers have adopted creative techniques to minimize the problem of social desirability bias, using experimental survey designs to avoid direct group comparisons. These methods build on the split-ballot survey design, in which randomly chosen subsamples of a survey are primed with one of several variants of a survey question to assess responses to a particular group or condition (Sniderman and Grob 1996).[1] For instance, Schuman and Bobo (1988) used a split-ballot design in which half the sample was asked about objections to a Japanese family moving into their neighborhood, while the other half was asked about objections to a black family moving into their neighborhood. Had each respondent been asked about both a black and a Japanese family on the same survey, they may have biased their responses toward similar evaluations of the two groups, consistent with norms of equal treatment. Through statistical comparisons across the two groups, split-ballot studies are thought to produce valid population-level estimates of the importance of race for the question of interest while reducing concerns about social desirability bias that arise from direct racial comparisons. Experimental survey designs have clear advantages for the measurement of sensitive topics, and their results have indeed shown a greater incidence of prejudice than those from traditional survey designs (Schuman 1995; Schuman and Bobo 1988).

[1] As the name implies, split-ballot designs commonly use two experimental groups. More recent work has taken advantage of computer-assisted telephone interviewing to extend this design to include many more variations in survey questions, including variation across multiple dimensions (see Sniderman and Grob 1996 for a review).

We view social desirability bias, then, as a problem that has received substantial attention in the research literature, with some promising developments. Nevertheless, to our knowledge, no research has provided a behavioral validation of experimental survey results. We have little concrete evidence, therefore, with which to evaluate when and to what degree experimental survey measures are in fact accurate proxies for behavior.

A second problem in using attitudinal measures as proxy assessments for discriminatory behavior concerns the emphasis of this method on consciously held beliefs or feelings. With the use of such measures, subjects are typically prompted for their attitudes in ways that allow for some degree of conscious deliberation. However, a growing literature in psychology has documented the existence and influence of implicit attitudes toward stigmatized groups that may influence judgments and actions without conscious awareness (Devine 1989). The intrapsychic processes that promote discrimination are likely to be more strongly activated in the context of a live interaction than in the abstract context of a survey question (Fiske 2004). Discrimination resulting from these interaction-triggered implicit stereotypes would necessarily remain undetected in survey responses, even those using an experimental design.

Creating a similar problem, some measured forms of discrimination may be perceptible only in the context of direct interaction. Social psychological evidence suggests that whites commonly experience heightened levels of social discomfort in the presence of blacks, at times leading to behaviors that are in effect discriminatory despite (accurately reported) nonprejudicial attitudes (Poskocil 1977). For instance, Word, Zanna, and Cooper (1974) show that white subjects conducting mock interviews with trained black applicants make more speech errors, ask fewer questions, and terminate interviews more quickly than with similar white applicants (see also Dovidio, Kawakami, and Gaertner 2002). Again, because these forms of discomfort are activated only by direct social contact—not by questions about hypothetical situations or prejudicial attitudes—these alternate psychological sources of discriminatory behaviors are unlikely to be captured by questions on survey instruments.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Finally, discriminatory action often is strongly influenced by situational factors, further reducing the extent of attitude–behavior correspondence. Complex decisions about where discrimination may be expressed are influenced by a combination of prevailing social norms as well as context-specific considerations (Merton 1949). In hiring, for instance, employers must balance their need for employees, the applicant pool, and other situational contingencies together with their taste for applicants along several dimensions. These situational factors can sometimes overwhelm the influence of prejudice in discriminatory action, resulting in a low correspondence between the two.

Indeed, it is notable, for example, that LaPiere's (1934) study found higher levels of racial bias apparent in the survey responses than in the field situation. Similar studies by Kutner, Wilkins, and Yarrow (1952) and Saenger and Gilbert (1950)—but focused on discrimination against blacks rather than the Chinese—report similarly counterintuitive results. These findings are especially remarkable in light of the contemporary literature on social desirability bias, which overwhelmingly assumes that survey reports will tend to underestimate the amount of discrimination that will occur. We believe the direction of the discrepancy between self-reports and behavior in these three studies most likely results from the importance of situational factors (Ajzen 1991). In the context of these studies, open discrimination likely would have involved some direct interpersonal confrontation. Unlike the decision not to call someone back for a job interview (a relatively passive form of discrimination), the refusal of service, or other more active forms of discriminatory treatment, can impose significant social costs.[2] In LaPiere's investigation, for example, the discriminator risked creating an uncomfortable interpersonal situation, possibly resulting in a scene. In certain cases, then, behaviors may be constrained in ways that verbal expressions are not, again leading to a lower correspondence between the two.

The historical evolution of strong norms against openly racist statements makes it less likely that contemporary field studies would find nearly as high a level of openly expressed prejudice as found in the aforementioned three studies (Schuman et al. 2001). And more recent, if indirect, comparisons of attitudes and behaviors usually have found stronger signs of racial discrimination in behaviors than in self-reports of behavioral intentions (Crosby, Bromley, and Saxe 1980). At the same time, it remains plausible that situational factors could still result in higher levels of stated than actual discrimination in certain cases, depending on the context and the attitudinal instrument.

The complexities involved in the relationship between attitudes and behaviors toward stigmatized groups emphasize the need for careful assessments of our measurement tools. The links between these measured attitudes and observed behaviors require systematic evaluation. Unfortunately, despite the frequent use of verbal expressions to draw conclusions about behaviors, very few studies directly calibrate survey responses with corresponding behavioral assessments.

Explicit Studies of Prejudice–Discrimination Correspondence

Whereas the sociological literature on the attitude–behavior relationship is small, the recent literature on the specific attitude–behavior case of prejudice and discrimination (in sociology) is virtually nonexistent.[3] Indeed, we turned instead to research in psychology for guidance in these matters. Social psychologist Susan Fiske (2004), in a recent, comprehensive meta-analysis, examine 54 studies containing empirical investigations of prejudice–discrimination correspondence. Consistent with the findings from the attitude–behavior literature more generally, Fiske finds wide variation in the relationship between prejudice and discrimination

[2] Laboratory studies have found that whites behave more aggressively toward blacks than toward other whites, but only when the consequences to the aggressor are low, such as when acting under conditions of anonymity, or in situations with limited possibility of retaliation or punishment (Crosby et al. 1980).

[3] Prejudice refers to negative judgments or opinions about a group (attitudes). Discrimination refers to unfavorable treatment directed toward members of a group (behavior).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

across studies, with correlations ranging from −.38 to .69, with a mean of .26. Her results thus support a general association between prejudice and discrimination, albeit at low average levels and with great variability across situations (see also Schutz and Six 1996).

The Fiske (2004) review, primarily featuring the work of psychologists, shows that sociologists have largely abandoned the study of prejudice–discrimination correspondence since the mid-1970s. Of the 10 articles in sociology journals included in the Fiske review, the most recent was published in 1973. This is not because sociologists have stopped using attitudinal measures and survey items to study discrimination against marginalized groups, as demonstrated by reviews such as those of Krysan (1999) and Schuman et al. (2001). Rather, sociologists have done little recent work to validate the assumption that these attitudinal measures are associated with discrimination. Krysan's (1999) review, for instance, notes the issue of attitude–behavior correspondence, but does not cite any recent studies on the topic. Instead, Krysan points to the similar trend directions for racial attitude items and corresponding behavioral indicators from unrelated samples and studies (Krysan 1999:139). Evidence of this sort does support a correspondence of attitude and behavior toward stigmatized groups, but only weakly so because a similar trend direction of indicators over time provides only very general evidence of meaningful correspondence.

In contrast to sociologists, among psychologists, the correspondence of attitudes and behavior toward stigmatized groups continues to be the subject of considerable research. Psychological research of this type has provided several important insights into the correspondence between different types of attitudes and behaviors, pointing to, for example, varying relationships between explicit/conscious attitudes, implicit/unconscious attitudes, and various forms of behavior (Dovidio et al. 2002). From a sociological standpoint, however, these studies have some important limitations, most notably those arising from a reliance on behavioral measures obtained in laboratory settings. For instance, the studies Fiske (2004) reviews use outcome behaviors such as ratings of perceived friendliness in interaction with a mock interviewer, subtle behavioral measures such as the number of blinks and the length of eye contact, or the results of role-playing situations. These outcomes often are far removed from the actual decisions made in their social contexts—to hire, to rent, or to move, to name a few—that are most relevant to understanding the behavioral processes that produce disadvantage among members of stigmatized groups.

For our purposes, the most relevant studies comparing prejudice and discrimination are those that assess these factors in realistic social settings, focusing on forms of discrimination that produce meaningful social disparities. Unfortunately, the three studies that fit this description each were conducted more than 50 years ago (Kutner et al. 1952; LaPiere 1934; Saenger and Gilbert 1950). We have very few means by which to assess the correspondence between contemporary racial attitudes and the incidence of discrimination.

Employer Attitudes and Hiring Decisions

The current study provides an opportunity to investigate these processes in a contemporary context. Bringing together a unique combination of data, we present a direct comparison of self-reported attitudes and corresponding behavior in the context of a real-world setting with important implications for inequality. The substantive focus of this study is on employers' willingness to hire blacks and/or ex-offenders for an entry-level position in their company. In both cases, the sensitive topics under investigation lead us to question the use of employer reports alone. By calibrating the estimates we received from surveys with behavioral measures from an experimental audit study, we are able to gain insight into the consistency between these two important indicators of group preference.

Measures of attitudes come in many forms, ranging from abstract statements of feelings (e.g., "I don't like members of group X") to more concrete statements of intended action (e.g., "I would not hire members of group X"). The latter, referred to as behavioral intentions, are considered the form of attitude that should most closely correspond to observed behavior, because of their conceptualization in terms of specific measurable action (Fishbein 1967; Fishbein and Ajzen 1975; Schuman and Johnson 1976). Thus a weak relation between behavioral intentions and behavior suggests an even weaker relation between the behavior and more

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

general attitudinal measures. In the current study, we rely on the behavioral intentions expressed by employers as an indicator of their attitudes about blacks and ex-offenders. Comparing what employers said they would do in a hypothetical hiring situation with what we observed them doing in a real hiring situation forms the basis of our current investigation.

METHODS

In the first stage of the study, employers' responses to job applicants were measured in real employment settings using an experimental audit methodology. Between June and December of 2001, matched pairs of young men (testers) were sent to apply for a total of 350 entry-level job openings in the Milwaukee metropolitan area.[4] The two white testers (one with a fictional criminal record and one without) applied for one set of randomly selected jobs (n = 150), and the two black testers (using profiles identical to those of the white pair) applied for a second set of jobs (n = 200).[5] The preferences of employers were measured based on the number of call-backs to each of the applicants, as registered by four independent voice mail boxes. Additional voice mail boxes were set up for calls to references listed on the testers' resumes.

[4] Jobs were randomly selected from ads placed in the Sunday classified section of the *Milwaukee Journal Sentinel* and on Jobnet, a state-sponsored Internet job site. Entry-levels jobs are defined as those requiring no more than a high school degree and limited work experience. Testers were 23-year-old college students from Milwaukee chosen for their effective styles of self-presentation and for their comparability in terms of physical and interpersonal attributes.

[5] The tester pair rotated which member presented himself as the ex-offender for each successive week of employment searches, such that each tester served in the criminal record condition for an equal number of cases. By varying which member of the pair presented himself as having a criminal record, unobserved differences within the pairs of applicants were effectively controlled. The criminal record used in all cases was a felony drug conviction. Although the more general term "ex-offenders" is used in reference to this group, it is important to note that the relationship between attitudes and behaviors may differ for individuals convicted of different crimes.

For a more detailed discussion of the research design, see Pager (2003).

The findings of the audit showed large and significant effects of both race and criminal record on employment opportunities. Call-backs were received by 34 percent of whites with no criminal record, 17 percent of whites with criminal records, 14 percent of blacks without criminal records, and 5 percent of blacks with criminal records (Pager 2003). Thus, overall, blacks and ex-offenders were one-half to one-third as likely to be considered by employers, with black ex-offenders suffering the greatest disadvantage.

The second stage of the study provided employers with the opportunity to express their hiring preferences verbally in the context of a telephone survey. Several months after completion of the audit study, each of the 350 employers was called by interviewers from the Michigan State Survey Research Center and asked to participate in a telephone survey about employers' hiring preferences and practices (see Pager [2002] for more detailed discussion of the survey instrument and results). Calls were directed to the person in charge of hiring for each establishment. The final survey sample included 199 respondents, representing a 58 percent response rate (Appendix A).

During the course of this survey, employers were read a vignette describing a job applicant with characteristics designed to match closely the profile of the testers in the audit study. Employers who had been audited by white testers were read a vignette in which the hypothetical applicant was white, and employers who had been audited by black testers were read a vignette in which the applicant was black. In this way, the survey design mirrored the split-ballot procedures used by Sniderman and Piazza (1993) and Schuman and Bobo (1988), avoiding direct racial comparisons within the same survey.

The wording of the vignette was as follows:

> Chad is a 23-year-old [black/white] male. He finished high school and has steady work experience in entry-level jobs. He has good references and interacts well with people. About a year ago, Chad was convicted of a drug felony and served 12 months in prison. Chad was released last month and is now looking for a job. How likely would you be to hire Chad for an entry-level opening in your company?

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Employers were asked to rate their likelihood of hiring this applicant with the following range of responses: very likely, somewhat likely, somewhat unlikely, and very unlikely.

The vignette presented in the survey was designed to correspond closely to the profile of the testers in the audit study. Chad, the hypothetical applicant, was presented with levels of education, experience, and personal qualifications similar to those on the resumes presented by the testers. The type of crime was identical, although the prison sentence in the vignette (12 months) was shorter than that reported in the audit study (18 months).[6] Thus the vignette aimed to measure employers' self-reports concerning how they would respond to such an applicant, whereas the audit measured how they actually did respond to an applicant with almost identical characteristics. The parallel scenarios of the vignette and audit should maximize the correspondence between the two measures (Schuman and Johnson 1976).

In the current study, the primary outcome of interest represents the employers' willingness to hire an applicant depending on his race and criminal background. As described earlier, in the survey, employers were asked to report how likely they would be to hire the applicant described in the vignette. In the actual employment situations, by contrast, we measured the number of employers who responded positively to testers after they had submitted their application. In most cases, this simply involved the employer inviting the tester to come in for an interview, although in a few cases, the applicant was offered the job on the spot. As we later discuss, the behavioral indicator should thus provide a highly inclusive measure of "willingness to hire," given that a call-back represents only an initial step in the hiring process.

RESULTS

In the following section, we examine the relationship between the survey results and the audit study. Initially, we compare the *level* of willingness to hire blacks and ex-offenders indicated by the audit results and the survey. We then examine the *association* between the two measures, considering whether employers who indicated high willingness to hire ex-offenders in the survey called back testers in the criminal record condition at higher rates than those who indicated low willingness to hire in the survey. In each of these comparisons, we seek to assess the degree to which what employers say is accurately reflected in what they do.

Figure 1 presents the key results from both data sources. The first two columns represent the percentage of employers who reported that they would be "very likely" or "somewhat likely" to hire the hypothetical applicant, depending on whether he was presented as white or black. We include the "somewhat likely" group here to correspond to our behavioral measure, which is a call-back rather than an actual hire (see discussion below).

The second two columns represent results from the audit study, illustrating the percentage of call-backs received by each group. In the audit study, call-backs also can be considered a measure of "willingness to hire," given that this represents a first cut in the hiring process. Although a call-back is by no means a guarantee of employment, given that employers typically call back several applicants before selecting their preferred hire, it does indicate a favorable initial review.

The results of the two outcomes, however, are anything but comparable. As Figure 1 shows, employers reported a far greater willingness to hire drug offenders in the survey than was found in the audit. In the survey, more than 60 percent of the employers said they were somewhat or very likely to hire a drug offender irrespective of the applicant's race. In the audit, by contrast, only 17 percent of white and 5 percent of black applicants with drug felonies actually received a call-back.[7]

The disparities apparent in these results are extremely consequential for our understanding of the social world. In the survey data, employ-

[6] The length of sentence was varied moderately between survey and audit to avoid arousing suspicion.

[7] Call-back rates include all the employers from the audit study, even those who did not complete the survey. Including only respondents captured in both samples produces even more disparate results. Call-back rates for white and black ex-offenders in the overlapping sample are 14 and 3 percent, respectively, demonstrating an even greater distance from the survey results.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
From callbacks to labor markets - In-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

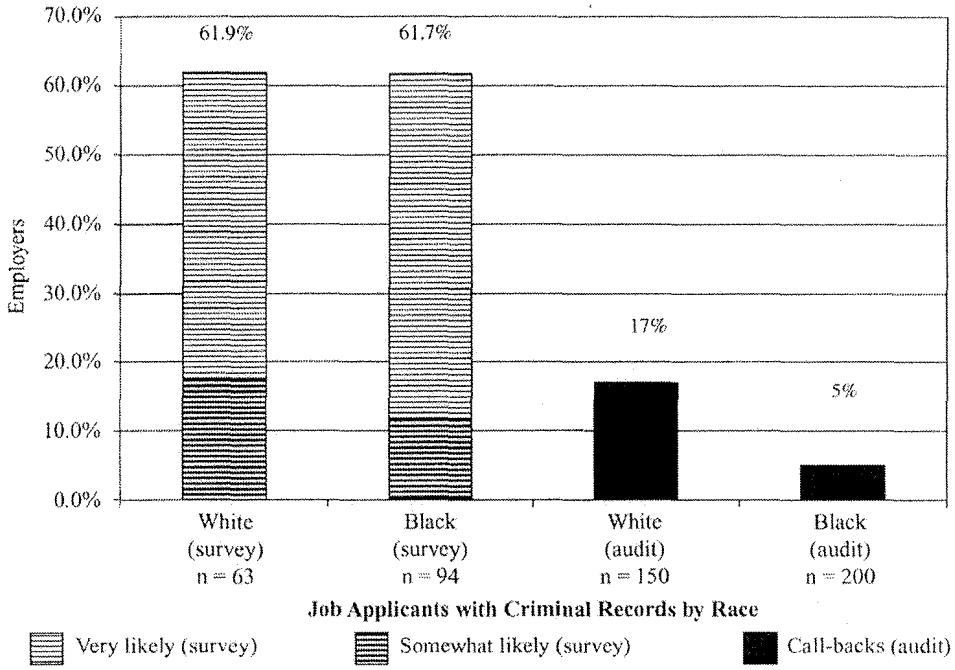


Figure 1. Expressed Willingness to Hire a Drug Offender According to Employer Survey and Audit

Note: Survey results include employers who said they were "very likely" or "somewhat likely" to hire the hypothetical applicant (with "very" at bottom of columns). Audit results represent the percentage of call-backs for each group. Differences between within-race comparisons of survey and audit results are significant on the basis of a two-sample test of proportions ($p < .05$)

ers' responses present a view of openness to blacks or applicants with drug felonies that is far greater than the reality measured in actual hiring situations. Accepting the survey results as an accurate indicator of the opportunities available to blacks and ex-offenders would grossly understate the barriers to employment they face.

Although the results of this initial comparison are compelling, there remain several possible objections to equating the findings from our survey measure with those from the audit study. First, collapsing the categories of "very" and "somewhat" into one category may artificially exaggerate the distance between survey and audit results. Indeed, if we look only at the "very likely" category, the discrepancy is far less striking. There is a literature on the meaning of vague quantifiers that attempts to offer greater precision to our understanding of these terms (Pace and Friedlander, 1982; Schaeffer, 1991). Lichtenstein and Newman (1967), for example, report that respondents assigned a mean probability of .87 to the phrase "very likely" and a mean probability of .59 to the phrase "somewhat likely." Whatever the exact probabilities to which these terms correspond, this literature indicates that such phrases imply a greater likelihood of hire than not. Remember that employers with greater reservations about the applicant also had the option of "somewhat unlikely" to indicate their ambivalence.

Whereas the survey asked employers to rate their likelihood of *hiring* the applicant, the audit merely measured whether the applicant was invited back for an interview or not. Although a call-back may represent a necessary condition to the decision to hire, it is by no means sufficient. In fact, according to the survey results, employers reported interviewing an average of eight applicants for the last noncollege job they had filled. Furthermore, employers on average reported interviewing 55 percent of the applicants that applied (Pager, 2002). Although these self-reported estimates may be inflated, they provide some evidence that the interview stage is far from synonymous with a hire. Rather, a call-back may in fact represent a fairly low bar of approval.

Thus, despite the different metrics on which our measures are based, we believe they provide roughly comparable indications of interest in the

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

applicant, corresponding to a moderately favorable review. In the results presented later, we provide an analysis of individual-level correlations that should be unaffected by these concerns.

A second possible objection to this comparison is that the very framing of the vignette item may artificially exaggerate the difference between survey and audit results. When considering a hypothetical applicant, employers do not have to take into account alternative possibilities among the applicant pool. Thus the hypothetical applicant may exceed the minimum threshold for acceptability even if in actuality there tend to be other applicants who are better qualified. By contrast, the tester in the audit study is competing with a pool of real applicants of varying quality. To the extent that real applicants provide better qualifications than does the tester's profile, the tester will receive few call-backs for reasons unrelated to race or criminal record.

An alternative way of presenting the information that addresses this concern is to calculate the likelihood that a tester with a criminal record will receive a call-back *relative to* a white tester without a criminal record. White testers without criminal records in this case represent a kind of baseline, presenting a given set of qualifications common among all testers, but without the handicaps of minority status or a criminal record. Employers who made call-backs to white testers without criminal records signaled that this level of education and experience was sufficiently desirable to make the first cut. Relative to this baseline, we can assess the proportion of blacks and whites with criminal records who received call-backs, thereby reducing the effect of employer nonresponses attributable to extraneous factors.[8]

Figure 2 displays the results of this procedure, comparing the likelihood of hire based on the survey and audit results with audit results recalculated as a ratio of the percentage of testers in the offender condition who received call-backs to the percentage of white testers with identical qualifications but no criminal background who received call-backs. Overall, 34 percent of white applicants with no criminal records, and with the given set of human capital characteristics presented by all testers, received call-backs. This group serves as our baseline (denominator) for calculating the relative call-back rates for the other groups. Only 17 percent of white testers with identical characteristics plus a criminal record received call-backs, indicating that white testers with a criminal record were 50 percent as likely to receive call-backs as those without a criminal record (Figure 2). Black ex-offenders were the least likely to continue in the employment process—only 5 percent received call-backs—indicating that they were just less than 15 percent as likely to receive a call-back as a similar white tester without a criminal record.

The differences between self-reports and behaviors in this comparison, although smaller, remain consistent when call-back frequency is judged relative to that for white nonoffenders. In the case of white ex-offenders, the distance between the survey and audit results has narrowed substantially, although it remains marginally significant statistically. The case for black applicants, on the other hand, maintains a clear and dramatic difference. Even relative to contemporaneous call-back rates for white testers, the call-back rate for black ex-offenders (14.7) remains far short of the survey estimates of hiring likelihoods (61.7). For black ex-offenders, the survey and audit measures provide dramatically different indications of willingness to hire.

Whatever measure is used, two main findings remain clear: First, whereas the survey responses present a rather benign view of the employment barriers facing ex-offenders, the audit results tell a very different story. Employers indicate a high level of willingness to hire drug offenders, but in actual employment situations,

[8] Similarly, we can consider the proportion of employers who reported that they were likely to hire an ex-offender (61.9 or 61.7 percent) as relative to an implicit baseline of 100 percent for a hypothetical applicant similar to the one described in the vignette, but without a criminal record. To the extent that some employers would report it not likely that they would hire such an applicant (if, for example, they hire only applicants with college experience), the ratio of the self-reported likelihoods of hiring an ex-offender relative to a nonoffender would be even larger, thus rendering the contrast between the self-reports and behavioral outcomes even greater.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

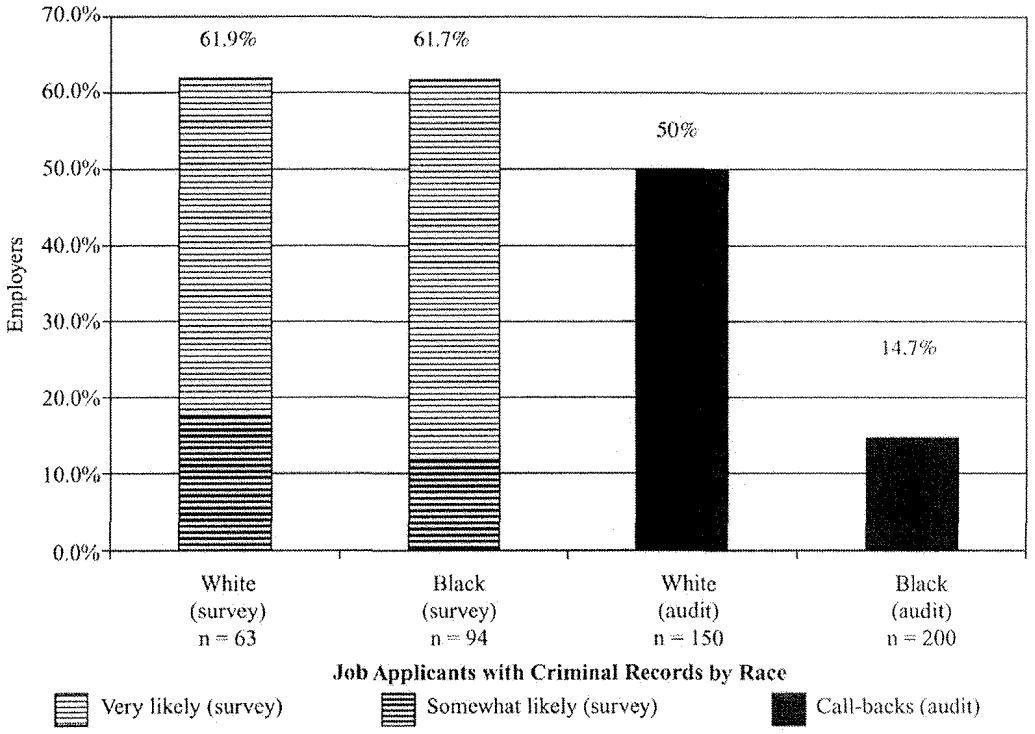


Figure 2. Expressed Willingness to Hire a Drug Offender According to Employer Survey and Recalibrated Audit

Note: Survey results include employers who said they were "very likely" or "somewhat likely" to hire the hypothetical applicant (with "very" at bottom of columns). Audit results represent the ratio of the percentage of call-backs for each group to the percentage of call-backs for white nonoffenders. Differences in within-race comparisons of survey and audit results are marginally significant for white applicants ($p < .06$) and significant for black applicants ($p < .05$) on the basis of a two-sample test of proportions.

they are less than half as likely even to call back such applicants relative to those without criminal records. This result underscores the importance of using great caution in relying on employers' self-reports as an accurate reflection of behavior.

Second, the degree to which race is a factor in hiring decisions is virtually undetectable in the survey results, in sharp contrast to what we find in the audit study. Table 1 shows the relative risk of receiving a call-back for white and black applicants in the survey and audit. In the survey, although separate employers were asked the vignette in which the hypothetical applicant was white or black, the estimates of hiring likelihoods for both applicants were virtually identical.[9] By contrast, actual behavioral measures in the audit show that white ex-offenders are more than three times as likely to receive consideration from employers as black ex-offenders.[10] These results suggest that employer surveys, even those with split-ballot designs, do not always provide an effective way to gauge the degree to which sensitive characteristics such as race affect actual employment opportunities. Later, we discuss the methodological and theoretical processes that might account for these differences.

Finally, we turn to the issue of individual-level consistency between survey reports and audit results. Even if the *levels* of openness to hiring ex-offenders are inconsistent between survey and audit, it remains possible that a *correlation* exists between the two: Employers who indicate

[9] Even if we were to restrict our attention to the "very likely" category, the black–white ratio (1.5) still would vastly understate the degree of racial disparity apparent in the audit.

[10] This result is calculated as a ratio of the call-back rate for white drug offenders (25/150) relative to the call-back rate for black drug offenders (10/200).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Table 1. Comparison of the Influence of Race on Hiring: Audit versus Survey

Relative Willingness to Hire	Survey	Audit
White vs. Black (White/Black)	1.00	3.33
95% CI	(.78, 1.29)	(1.65, 6.73)

Note: The difference in white/black ratios between the survey and audit is statistically significant at $p < .001$ (Mantel–Haenszel test, $\chi^2 = 12.01$, 1 df).

Table 2. Individual-level Consistency between Employers' Self-reports and Behavioral Outcomes

	Audit Results (for Testers Presenting Drug Felony)	
Survey Results	No Call-Back	Call-Back
Likely to Hire Drug Offender		
No	56 (93.3 %)	4 (6.7 %)
Yes	89 (92.7 %)	7 (7.3 %)
Difference of Percentages (95% CI)		0.6% (–8.6%, 8.8%)
Correlation (Kendall's Tau-b) (95% CI)		.012 (–.143, .167)

Note: This table includes all employers who responded to the survey. Call-backs in the right column above represent calls to the tester in the criminal record condition only.

willingness to hire ex-offenders may be more likely to hire an ex-offender than employers who do not indicate such willingness, even if the overall openness to hiring ex-offenders is strongly overstated in the survey results. This final analysis allows us to compare the survey responses with the audit outcomes at an individual rather than an aggregate level. The results of this cross-tabulation are presented in Table 2.[11] Consistent with the results reported earlier, we find that the survey responses have very little connection to the actual behaviors exhibited by these employers.

Among those who reported a favorable likelihood of hiring an applicant with a prior drug conviction in the survey, 7.3 percent made calls to the tester with the criminal record in the audit study, relative to 6.7 percent of those expressing an unfavorable likelihood. This difference is in the expected direction, but is only slightly greater than zero (0.6 percent), and far too small to reach statistical significance. Likewise, correlational measures for ordinal data, such as Kendall's Tau-b, show nearly zero association between survey and audit.[12] Considering the possibility that our relatively small sample size

[11] The percentage of call-backs is lower than for the aggregate comparisons because of survey nonresponse. A full breakdown of survey responses (including all four survey response categories) by audit results is presented in Appendix B. Given the small sample sizes in this final comparison, a further breakdown by race of the tester would be impossible. Analyses, therefore, include all call-backs to testers in the criminal record condition regardless of race.

[12] The individual-level comparison further allows us to reconsider the practice of combining "very" and "somewhat" likely responses into a single category. To investigate this question, we recalculated Kendall's Tau-b between the audit call-backs and the vignette question using the four survey answers, with separate "very" and "somewhat" categories. Instead of producing stronger results, this coefficient is not statistically significant, and the negative sign is opposite that expected (Tau-b = –.0391).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

limits the reliability of these estimates, we calculate confidence intervals allowing us to assess the potential relationship that would obtain if we had used a larger sample. A 95 percent confidence interval for the difference in percentages includes a range from –8.6 to 8.8 percent, indicating that we are 95 percent confident that employers who indicate "yes" are no more than 8.8 percentage points more likely to make a call-back than employers who indicate "no."[13] A difference of 8.8 percentage points for making a call-back still is a fairly low level of correlational consistency.[14] We can thus be fairly confident that, given this pattern of results, even a much larger sample would be unlikely to produce a substantial relationship between survey and audit results.

These results cast strong doubt on the accuracy of survey data for indicating relative likelihoods of hiring. Individuals who report a higher likelihood of hiring an ex-offender are only trivially more likely to do so. Confirming the aggregate findings described earlier, the individual-level associations presented here appear to be no better at establishing a relationship between attitudes and behaviors.

Nevertheless, several limitations of this analysis must temper its conclusions. In the following discussion, we consider possible threats to the validity of our findings caused by measurement error or study design. In the first case, we consider the possibility for error in the survey or audit results, either of which could lead to a weakened correlation between the two.

In the case of the survey, the most plausible source of measurement error arises from those cases in which the survey respondent was different from the individual who reviewed the testers' applications. To the extent that hiring practices vary within firms depending on the individual manager or human resource officer, the consistency between survey and audit results will be attenuated. Although recent evidence suggests that labor market discrimination typically operates at the level of the firm rather than the level of individual discriminatory actors,[15] this possibility remains a potential source of measurement error. Nevertheless, although within-firm heterogeneity may indeed affect the individual-level consistency in measures of attitudes and behavior, there should be no effect on the average level of support for hiring ex-offenders in the aggregate (as presented in Figures 1 and 2). In cases wherein respondents differ, there is little reason to believe that the hiring agent would be systematically more or less likely to consider hiring ex-offenders than would the survey respondent.[16]

In the audit study, measurement error also may pose a problem. Because each employer was visited only once, we have only one data point with which to assess their hiring tendencies. Given the many factors at play in any given hiring situation (e.g., the urgency with which the position must be filled, the number and quali-

[13] This confidence interval is calculated using the "plus 4" method of Agresti and Caffo (2000). The Agresti and Caffo method has the advantage of providing accurate (and slightly conservative) intervals even when the count of successes or failures is very small. By contrast, the methods in standard introductory statistics books usually require at least five successes and failures for each group.

[14] Similarly, the upper limit of the 95 percent confidence interval for Kendall's Tau-b suggests that the degree of correlational consistency in a large sample probably would be quite low (below .167). The confidence interval for Kendall's Tau-b is asymptotic, and thus should be regarded with more caution than the confidence interval for the proportion, which is computed by a method with good small-sample properties.

[15] Recent studies have found that firm-level variables such as the presence of a human resource apparatus, the use of applicant tests, and affirmative action policies have far more influence on the racial composition of a company than the individual characteristics of hiring managers or owners (Holzer 1996; Holzer and Neumark 2000).

[16] One might also question whether possible changes in the economic climate at the time of each measurement may be responsible for some of the disparity. The unemployment rate in Milwaukee averaged 4.8 percent during the time of the audit, whereas in the 2 months during which the survey was administered, it had risen to nearly 6 percent (Bureau of Labor Statistics 2002). Given that employers' openness to less desirable workers increases in the context of tight labor markets (Freeman and Rodgers 1999), we would rather expect more favorable responses from employers during the period of the audit study relative to the period of survey data collection.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

ty of other applicants), any single data point representing an employer's treatment of an ex-offender may be subject to the measurement error of circumstance.[17] Nevertheless, the almost complete absence of association between survey and audit measures leads us to question random error as a sufficient explanation. If the hiring process tends to be so complex as to defy any straightforward relationship between the abstract intentions of employers and their ultimate decisions, this would imply more than a simple problem of measurement. Indeed, we later discuss the complexity of hiring decisions as one of our hypothesized explanations for the discrepancy between outcomes.

More systematic forms of error can creep into an audit design if the experimental procedures are somehow compromised. Of primary concern are the potential biases that can be caused by the performance of testers, either because of poor matching or because of testers' self-fulfilling expectations (Heckman and Seligman 1993). Fortunately, a rather direct test of this hypothesis is possible. One would expect that if differences in testers' personalities or behavior shaped the outcomes (above and beyond any effects of race or criminal condition), we should see stronger results among those cases in which the testers had the opportunity to interact with employers. Applications submitted with little or no personal contact should be far less affected by such concerns. Analyzing the outcomes for these two kinds of tests, we find no evidence of tester bias. In fact, the main effects of both race and criminal background are substantially attenuated among those who did have personal contact with the employer (Pager 2003). This suggests that instead of exaggerating negative stereotypes, the appealing characteristics of these testers actually worked to reduce the measured effects, thus biasing the results in a direction consistent with the survey responses. It is unlikely then that these various sources of error can account for the significant disconnect between the survey and audit results.

A final limitation of the comparison provided earlier is its reliance on a single survey item. Flaws in the survey design or peculiarities of the question wording could lead to anomalous patterns of responses. Fortunately, an additional item was included in the survey that allows for similar comparisons to be drawn. Whereas the vignette most strongly paralleled the audit situation—including a match of the applicant's race—the second item also probed employers' willingness to consider hiring ex-offenders, in this case focusing on a generic applicant with a criminal record. The exact wording of this item was "Next, I am going to list several types of applicants. Please tell me if you would accept each type for the [most recent noncollege] position . . . an applicant who has a criminal record?" As described earlier, we compare those who answered that they "definitely" or "probably" will hire with those who answered "definitely" or "probably" not in relation to the audit outcomes.[18] The correlation between this measure and the audit results is again very small. The correlation coefficient from Kendall's Tau-b is .0003 (95 percent confidence interval, –.154 to .155). Again, our tests for significance in this case cannot reject the hypothesis of no relationship.

The various aforementioned limitations must certainly temper our conclusions. Nevertheless, the almost total lack of correlation between the survey and audit results is troubling. If these findings are an accurate assessment of the level of consistency for these and related measures, then studies that use similar survey items to draw conclusions about characteristics or circumstances associated with discrimination may come to strongly misleading conclusions.

On the basis of several methods for assessing the attitude–behavior relationship, all comparisons tell a similar story: it is difficult to get

[17] As in the case of the survey, measurement error in the audit results will attenuate individual-level associations, as presented in Table 2, but should not affect the aggregate comparisons presented in Figures 1 and 2.

[18] In the original survey, 25 percent of the respondents gave the response "it depends." As a conservative estimate, these respondents were treated as "willing to hire," producing a stronger correlation than when they are excluded from the analysis. In fact, the association of survey and audit becomes slightly negative when this category is excluded. (By contrast, the correlation coefficient from Kendall's Tau-b for this survey item and the original vignette item was .55).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.
803

an accurate picture of actual hiring outcomes based on responses to the employer survey used in this study. Employers generally express a greater likelihood of hiring applicants with criminal records, and a far greater likelihood in the case of black applicants, than we see in actuality. Furthermore, employers who indicate greater willingness to hire an ex-offender in response to a survey question seem to be only slightly more likely actually to offer an interview to such an applicant. Both in terms of making aggregate- and individual-level predictions, our evidence points to weak correspondence between survey results and actual hiring outcomes.

ATTITUDES AND BEHAVIORS: WHY DO THEY DIFFER?

Why might employers' survey responses present results so discrepant from their actual behavior? Several theoretical explanations could be used to account for this incongruity. In this section, we provide a discussion of these explanations, considering the range of underlying processes that may give rise to differing outcomes.

Social Desirability and Compensatory Estimation

As discussed earlier, efforts to measure attitudes about sensitive topics are complicated by the problems of social desirability bias. According to this perspective, respondents may conceal their true feelings about blacks or ex-offenders in answering survey questions. If this is the case, the discrepancy between self-reports and behaviors should be viewed as the difference between false and true measures of a respondent's attitudes. Although social desirability pressures certainly result in some distortion of survey results, we do not believe that this can account fully for the differences between expressed willingness to hire ex-offenders and the actual employment outcomes based on applicants' criminal record. In fact, at other points in the survey, respondents expressed strong opposition to considering applicants with criminal records other than drug felonies: nearly 70 percent of employers expressed an unwillingness to hire an applicant who had been convicted of a property crime, and more than 75 percent were self-described as unwilling to hire an applicant who had been convicted of a violent crime (Pager, 2002; see also Holzer, Raphael, and Stoll 2003).[19] It therefore seems unlikely that social (or legal) pressures to accept ex-offenders whitewashed employer responses. High levels of acceptance were reported only for the applicant described as a drug felon.

Social desirability bias may be a greater concern in the measurement of racial preferences from the survey results, which is where we find the largest disparities between expressed attitudes and observed behaviors. To preempt this concern in the current study, we used a split-ballot format in which each employer responded to only one hypothetical (black or white) candidate. It remains possible, however, that social desirability bias is a problem if, even in the absence of direct comparisons by race, employers are aware that the race of the hypothetical applicant has been specified and therefore make conscious or unconscious efforts to compensate verbally for any negative reactions they may have to a black applicant. If respondents do in fact suppress negative reactions to race-specific targets, even when no racial comparison is provided, this calls into question the effectiveness of split-ballot survey designs as a strategy for measuring underlying racial prejudice. Any self-reported attitude toward a black target may in fact be distorted by the respondent's own compensatory estimation procedure.

Abstract Versus Specific Considerations

A second possible explanation for the discrepancy between measured attitudes and behaviors in this study relates to differences in the criteria used to evaluate a hypothetical versus an actual job candidate. It is plausible that the affirmative responses of employers considering the acceptability of a hypothetical applicant

[19] Questions about specific crime types followed the presentation of the initial vignette described earlier. Employers were asked to report the likelihood of hiring Chad if, instead of a drug crime, he had been convicted of a property crime, such as burglary, or of a violent crime, such as assault. Response categories for property/violent offenders were "very likely" (10/7 percent), "somewhat likely" (21/17 percent), "somewhat unlikely" (32/29 percent), and "very unlikely" (37/47 percent).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

indicate their genuine willingness to consider hiring an applicant with a criminal record *in the abstract*. In these general terms, apart from the minority of employers who categorically reject all applicants with criminal records, a prior conviction is not typically grounds for immediate disqualification. Rather, if the applicant's overall characteristics exceed a minimum threshold of employability, the respondent is likely to indicate a willingness to hire.

By contrast, in actual employment situations, the applicant's characteristics are judged not only according to some minimum threshold, but also relative to the pool of available applicants, and to the specific requirements of a job. In this case, many more contingencies are at play, and the presence of a criminal record may become a salient criterion by which to weed out less-qualified applicants.[20] Even if the employer genuinely believes that she or he would hire the applicant described in the abstract vignette, when confronted with the situation in real life, the contingencies of the hiring process may render hypothetical scenarios irrelevant.

Recognizing this potential disconnect, we made efforts to calibrate the behavioral responses to a concrete indicator of employability at that place and time. In this case, the white nonoffender, our baseline in Figure 2, serves this function by providing assurance that this level of qualification was sufficient to elicit a callback during that particular hiring process. Despite this adjustment, the willingness to hire expressed in the survey appears to be much higher than in the audit.

Even if differences between the exact vignette and the audit situations can explain some of the discrepancies between survey and audit measures of overall willingness to hire, this explanation cannot account for the considerable difference in race effects detected by these comparisons. In the survey, black and white applicants appear equally likely to receive offers, whereas in the audit, there is a large gap in favor of white applicants. An investigator using these survey data alone would be strongly misled about the role of race in shaping actual hiring decisions.

The Intensity Of Priming

A third perspective on the discrepancy between self-reports and behaviors proposes that the priming of characteristics during a phone interview may not elicit the same intensity of response as the in-person presentation of the same characteristics. Hearing a description of a hypothetical black ex-offender is quite different from seeing a young black man approach one's business in search of employment. The live interaction may trigger feelings of fear, anxiety, or threat in ways that a recited vignette does not (Poskocil 1977). These feelings may then influence employment decisions in ways that cannot be fully replicated in hypothetical scenarios in surveys.

Similarly, social psychological evidence suggests that racial stereotypes exert many of their effects indirectly, by coloring the evaluation of ambiguous information (Darley and Gross 1983). When employers are evaluating applicants, for example, an energetic, outgoing, young white applicant may be viewed as motivated and eager to work, whereas a similarly energetic, outgoing, young black male may be seen as a hustler or a "player." Even relatively straightforward cues can be interpreted in vastly different ways, depending on the context of the situation or the characteristics of the actor (Sagar and Schofield 1980). Again, these sorts of distortions would most likely operate during in-person evaluations. The vignette, by contrast, leaves less room for distorted interpretations, because according to the explicit description, the hypothetical candidate "has good references and interacts well with people." This class of explanations suggests that discriminatory behavior in the employment of ex-offenders, especially African-Americans, may have a basis in sources other than consciously believed attitudes toward these groups.

[20] Indeed, in response to the second more general survey question discussed earlier, when employers were asked about their willingness to hire a generic applicant with a criminal record (with no additional information provided), a large fraction of employers refused the forced-choice response categories, insisting instead that "it depends"—on the crime, on the length of time since the conviction, on the type of job, and numerous other considerations. In this case, then, the employers themselves acknowledged that any estimation given in the abstract would have very little bearing on how they might respond in making an actual hiring decision.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

In the article discussed earlier, LaPiere (1934) reinforces the view that surveys may elicit a different set of considerations than do concrete experiences. According to LaPiere, survey responses constitute "verbalized reaction[s] to a symbolic situation," or reactions to a highly abstracted representation of reality (p. 231). According to this viewpoint, survey responses do tell us something meaningful about the attitudes of respondents, but we have no way of anticipating the degree to which these expressed attitudes will be reflected in any particular set of behaviors. Certainly, it is difficult to anticipate how any individual, including oneself, may react to a situation previously encountered only in hypothetical terms. In the case of hiring decisions about individuals with stigmatizing characteristics, our results suggest that very little can be implied from these self-reports of employers for the accurate prediction of employment outcomes.

It is not possible using the current data to demonstrate conclusively which underlying process may have generated the observed discrepancies. In fact, it is highly plausible that more than one process may have been at work simultaneously. What these results do demonstrate, however, is that employers' expressed willingness in the survey taps into a set of processes very different from those measured through our behavioral study. Although these processes may be related to a common underlying disposition, the correspondence between the two can be quite weak. It is important that researchers recognize these limitations before drawing inferences about behavior from the self-reports of survey respondents.

Rethinking the Role of Attitudes

What can we conclude from these results regarding the usefulness of data on attitudes? Should we disregard all employers' self-reports? Certainly, it would be premature to advise such a radical stance. In fact, despite the large discrepancies between the self-reports and behaviors measured in the current study, we believe that survey results remain useful, even if they cannot be viewed as an alternative procedure to the measurement of actual discrimination.

Even in cases in which expressed attitudes have little relationship to measured discrimination, survey data can nevertheless tell us something useful about how employers *think* about important hiring issues. Responses to the survey suggest, for instance, that many employers who discriminate against blacks do not necessarily do so because of a principled belief that black employees should not be hired. In fact, we think it likely that many employers genuinely believe their own responses to surveys, professing the value of equal opportunity, while simultaneously justifying their behavior in hiring situations on grounds other than race (e.g., assumptions about the family/social/educational backgrounds of black applicants; see Kirschenman and Neckerman 1991). In this case, the difference between employers' self-reports and their actual behavior represents a meaningful discrepancy between two legitimate realities. The resolution of these differences represents an important focus of sociological investigation in its own right. Although low correlations between attitudes and associated behaviors often are viewed as a purely methodological test of survey questions, in many cases, these discrepancies actually may provide clues toward a better substantive understanding of the cognitive-emotional basis for action.

Furthermore, it remains possible that survey research may provide a better proxy for behavior in situations that are less complex and subject to fewer contextual influences than hiring. Action in any real social situation is the result of many factors other than the actor's attitude toward the object, including norms, perceived consequences of the action, and implicit or unconscious attitudes toward the object. The many complex influences on hiring decisions make these situations exactly the sort for which survey measures are least likely to be an effective substitute. Indeed, the three "classic" studies that found very weak associations between expressed behavioral intentions and behaviors all were studies of discrimination in social situations (Kutner et al. 1952; LaPiere 1934; Saenger and Gilbert 1950). We believe it possible that survey responses may provide a much more effective proxy for behavior in other contexts, such as those that involve voting (Traugott and Katosh 1979), signing of a petition (Brannon et al. 1973), or patterns of consumer behavior (Day et al. 1991), in which the link between behavioral intentions and actual behavior is less subject to contextual influences apart from the respondent's attitude or intention.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Finally, we have focused on only a few of the many survey techniques that have been developed to measure prejudice and discrimination. Though our measure of behavioral intentions was designed to offer the closest match to the audit context, it remains possible that other more abstracted measures of racial bias may in fact correlate more closely with measures of discrimination. There is an extensive literature that attempts to investigate modern or subtle forms of racial attitudes using survey questions (National Research Council 2004, chapter 8), and certain of these alternative approaches could prove more effective at capturing behavioral outcomes than what we found in this study.

Three sociological approaches that we believe to be especially promising place their respective emphasis on stereotype measurement, past behavior, and in-depth interviews. The first of these, group stereotype measurement, has a long history in the social sciences, with research demonstrating a persistence of racialized attributions across numerous dimensions (Devine and Elliot 1995). As a recent example, researchers have developed a series of scales measuring respondents' images of different racial groups along a wide range of social and psychological characteristics (Bobo and Kluegel 1997; Smith 1991). Survey techniques such as these have shown respondents to rate blacks as worse or inferior relative to other groups on dimensions such as criminality and intelligence, suggesting that traditional measures of racial attitudes may be missing a great deal of underlying racial bias.

A second approach, used extensively by Harry Holzer and colleagues, asks employers to focus on the last worker hired, thereby grounding responses in a concrete recent experience (Holzer 1996; Holzer, Raphael, and Stoll 2003). By focusing on a completed action, Holzer is able to avoid the ambiguities of hypotheticals or general statements; and by focusing on actual outcomes, he is able to document "revealed preferences" rather than expressed ones. Likewise, Holzer's format calls for race to be assessed only as one incidental characteristic in a larger series of questions concerning the recent employee, thereby reducing the social desirability bias often triggered when the subject of race is highlighted. Whereas recall or motivational biases emerge as concerns in the reporting of prior experiences (Bradburn, Rips, and Shevell 1987), this particular approach focuses on a well-defined event that leaves less room for error-prone estimation.

Finally, some data suggest that in-depth in-person interviews may be more effective in eliciting candid discussions about sensitive hiring issues than other modes of interviews (Kirschenman and Neckerman 1991; Moss and Tilly 2003; Wilson 1996). In-depth interviews offer the opportunity for respondents to discuss the complexities and, at times, the inconsistencies in their views of various groups, thereby going beyond the more generalized assessments expressed in traditional survey items. Likewise, the opportunity for rapport-building in the in-person interview context may reduce social desirability pressures, making respondents feel at greater ease in expressing counter-normative beliefs.[21]

Although each of these represents a promising approach, our results caution against the unreflective assumption that the results of any method are necessarily good proxies for behavior. None of these techniques has yet been subjected to direct assessments by comparison of their responses with corresponding behavioral measures, a step we view as necessary for understanding the relation of these measures to behavioral outcomes. LaPiere's (1934) warning, that hypothetical scenarios often cannot convey the experience of concrete situations, deserves to be taken more seriously by current generations of survey and interview researchers.

CONCLUSION

LaPiere (1934) showed a striking inconsistency in the way hotel and restaurant proprietors reacted to Chinese customers in person, as compared with how they responded in surveys. The current study notes a similar discrepancy between employers' self-reported likelihood of hiring a particular applicant and their actual hiring behaviors when faced with a nearly identical candidate. We found an especially large and robust disparity between the reported likelihood of employers hiring black ex-offenders and actual rates of hiring. The low correlation between expressed and observed hiring out-

[21] For a critique, see National Research Council (2004), p. 175.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

comes presents an epistemological worry: our assessments of the degree of disadvantage faced by black ex-offenders would be substantially underestimated on the basis of the survey results alone. Moreover, we found little correlation between greater expressed likelihood of hiring ex-offenders in the survey and actual increased rates of call-backs for ex-offenders in real hiring situations. Given that most research on hiring preferences and practices comes from the self-reports of employers themselves (Downing 1984; Holzer 1996; Husley 1990; Jensen and Giegold 1976; Waldinger and Lichter 2003; Wilson 1996), these results indeed have serious implications.

In terms of the methods used to measure discrimination, these findings suggest that sociologists may need to reevaluate what is learned from studies that use vignettes of hypothetical situations to study behaviors toward stigmatized groups. Although we believe that these vignette studies often do tell us about respondents' abstract beliefs, in some cases these beliefs may have relatively little influence on the behavior of interest. Feelings and evaluations in a concrete social situation may be very different from those in the abstracted situation of the survey, but the two often are treated as nearly identical. An important next step in evaluating the contribution of survey measures for understanding behaviors of interest is to relate these items to actual behavior.

More broadly, these results suggest the limits of survey questions alone for understanding the changing nature of racial inequality. Survey questions indicating a liberalizing of racial attitudes among white Americans have been cited widely as evidence supporting the declining significance of race in American society. But if the items analyzed in this study have any bearing on survey responses more generally, we have reason to question that changing public opinion on matters of race has any necessary correspondence to the incidence of discrimination. Rather, our results support the perspective that there has been a growing gap between the principled statements and beliefs of white Americans in favor of racial equality and their concrete actions.[22] Survey questions provide one important perspective on American race relations, but they must be combined with other information for a complete picture.

Fortunately, methods to improve our understanding of the prejudice–discrimination relationship are readily available and feasible for even small groups of researchers. The comparison of survey measures and behavioral indicators does not require an unprecedented level of resources. Even for pedagogical purposes, sociology teachers could readily incorporate the dual design within a two-semester timetable. Whereas audit studies of labor markets can be quite involved, numerous other everyday social settings provide countless opportunities for implementing small-scale experiments: searching for an apartment, shopping, hailing a taxi, passing security checkpoints, and the like (for an example of a classroom application of the audit methodology, see Massey and Lundy 2001). Moreover, the investigation of prejudice and discrimination could be usefully applied to many other groups: Asians, Latinos, Muslims, women with criminal records, gays and lesbians, to name just a few.

For creative sociology teachers, then, a single class could readily achieve a paired audit study and telephone survey with sufficient sample sizes for meaningful comparisons. Both substantively and methodologically, the pairing of survey and audit research can yield important insights for the study of contemporary discrimination.

It is not the case that employers' thoughts and beliefs can tell us nothing about important employment issues. In fact, in many cases, surveys and other methods of eliciting employer opinions can provide useful information about attitudes and beliefs. In other cases, surveys may provide a very close reflection of actual employer behaviors. What this research emphasizes, however, is the importance of testing one's assumptions and providing external validation of key results. In the case of employers' behavior with respect to hiring black ex-offenders, the survey results presented here are far off base. The correspondence between self-reports and

[22] This perspective has also found support from certain well-designed surveys that manage to capture respondents' contradictory or competing beliefs (Bobo, Kluegel, and Smith 1997; Kinder and Sears 1981).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

behaviors with respect to other important hiring outcomes has yet to be established.

Devah Pager *is an Assistant Professor of Sociology at Princeton University. Her research focuses on institutions affecting racial stratification, including education, labor markets, and the criminal justice system. Continuing her work using experimental field methods, she currently is investigating discrimination in low-wage labor markets in New York City (with Bruce Western). Her recent publications include "The Mark of a Criminal Record," published in the American Journal of Sociology (March 2003), and "The Structure of Disadvantage: Individual and Occupational Determinants of the Black–White Wage Gap," published in the American Sociological Review (August 2001) (coauthored with Eric Grodsky).*

Lincoln Quillian *is Associate Professor of Sociology at Northwestern University and the University of Wisconsin–Madison. His research focuses on inequality, racial attitudes, and urban population distribution. His current research projects include studies investigating the role of stereotypically biased evaluations in the creation of racial inequality and the consequences of spatial income segregation for social stratification.*

APPENDIX A

Survey Design and Implementation

The baseline survey instrument was developed by Harry Holzer and his colleagues.[23] It includes questions about the company such as size, industry, employee turnover, and racial composition; questions about hiring procedures such as the use of interviews, personality or aptitude tests, and background checks; questions about the last worker hired for a position not requiring a college degree including age, race, and sex of the worker, recruitment method, wage, and promotion opportunities; and questions concerning the employer's attitudes about various kinds of applicants including welfare recipients, applicants with long spells out of the labor market, unstable work histories, or criminal records. In addition, several survey items were added to mirror the audit study more closely (as described earlier).

The survey was administered by the Michigan State Survey Center. The final survey sample included 199 respondents, representing a 58 percent response rate. Response rates were calculated according to the basic formula $(I + P)/(I + P + R)$, where I equals the number of completed interviews, P equals the number of partial interviews, and R represents the number of refused eligible numbers (Groves and Lyberg, 1988). Between the time of the audit and that of the survey, two companies had declared bankruptcy, and an additional two had nonfunctioning numbers. These firms were dropped from the survey sample and excluded from the denominator for calculations of response rates.

Typical response rates for academic telephone surveys range from 50 to 80 percent. The current survey falls toward the lower end of the range of acceptable response rates as the result of several possible factors. Response rates for surveys of top management and organizational representatives typically lag behind those of employees or the general population (Baruch 1999). Likewise, there has been increasing resistance of businesses to participation in surveys, given the proliferation of market research firms and academics seeking employer participation for the growing number of studies involving businesses (Remington 1992). There has been a notable downward trajectory in the response rates from business surveys over the past 25 years (Baruch 1999; Cox et al. 1995), with increasing numbers of refusals citing that participation was against company policy (Fenton-O'Creevy 1996, cited in Baruch 1999). Even among the general population, Curtin, Presser, and Singer (2000) reported that the number of calls required to complete an average interview and the proportion of interviews requiring refusal conversion doubled between 1979 and 1996. The inundation of telemarketers (and, to a lesser extent, survey research) matched by the technological advances of caller-ID and privacy managers has made it increasingly difficult to recruit survey respondents for academic research (Remington 1992).

To assess the possible bias that may result from selective participation, two comparison tests were made. The first test compared basic characteristics of employers who responded to

[23] The first version of this survey was developed for the Multi-City Study of Urban Inequality Employer Survey (Holzer 1996). Holzer, Stoll, and Raphael (2002) later modified the initial instrument to focus more closely on applicants with criminal records. The instrument used for the current study was further modified to reflect the priorities of this research project.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

the survey with the characteristics of those eligible for participation but refused. On the basis of industry, location, and call-back rates, the two groups were very similar, although some differences in occupational distribution were apparent: employers for restaurant jobs were the most likely to respond to the survey, whereas employers for laborer or service positions were the least likely. This difference probably has to do with the accessibility of employers in locally run restaurants, as compared with those in decentralized factories, warehouses, or companies. In an effort to account for this overrepresentation, key outcomes are recalculated using weights to achieve the sample distribution of the audit study (available upon request). A reweighting of the survey sample to match the distribution of the audit sample produced only a slight change in the mean likelihood (from .62 to .60). It is unlikely, therefore, that differential response rates of employers across industries have any effect on the survey outcomes or on the differences between survey responses and observed behavior.

Even without these adjustments, however, the distribution of responses on key attitude items closely match that for a previous sample of Milwaukee employers. In a second test of sample bias, basic employer characteristics from the current sample were compared with an identical set of questions asked of a more representative sample of Milwaukee employers in 1999 (Holzer and Stoll 2001). Although the earlier Milwaukee survey included a broader geographic area and oversampled large firms, the general attitudes expressed by employers in both samples were strikingly similar (Table A1). The consistency of these findings provides some reassurance that the current sample can serve as a useful gauge for the priorities and concerns of employers in the broader Milwaukee metropolitan area.

Table A1. Comparison of Employer Attitudes and Characteristics across Two Milwaukee Surveys

Variable	Pager 2002	Holzer and Stoll 2001
Employees (n)	66.95	180.47
Vacancies (n)	4.48	7.79
Minority-owned Companies (%)	8.40	8.41
Unionized Employees (%)	9.30	15.19
Industry		
Manufacturing (%)	12.43	20.00
Retail trade (%)	49.72	21.00
Services (%)	21.47	39.00
Other industry (%)	16.38	20.00
Hire Welfare Recipient		
Definitely/probably would (%)	97.40	96.62
Definitely/probably not (%)	2.60	3.37
Hire Applicant with GED		
Definitely/probably would (%)	98.80	97.23
Definitely/probably not (%)	1.20	2.77
Hire Applicant with Criminal Record		
Definitely/probably would (%)	49.20	49.20
Definitely/probably not (%)	50.80	50.80
Hire Applicant Unemployed >1 year		
Definitely/probably would (%)	70.90	80.15
Definitely/probably not (%)	29.00	19.86
Hire Applicant with Unstable Work History		
Definitely/probably would (%)	60.50	67.49
Definitely/probably not (%)	39.50	32.51

Note: GED = General Education Diploma.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

APPENDIX B

Table B1. Comparison of Employers' Self-Reports and Behavioral Outcomes for Overlapping Sample

	Audit Results	
Survey Results	No Call-Back	Call-Back
Likely to Hire Drug Offender		
Very Unlikely	31 (96.9 %)	1 (3.1 %)
Somewhat Unlikely	25 (88.3 %)	3 (10.7 %)
Somewhat Likely	69 (93.2 %)	5 (6.8 %)
Very Likely	20 (90.9 %)	2 (9.1 %)

Note: N = 199 respondents. Data shown as number of respondents with percent in parentheses.

REFERENCES

Agresti, Alan and Brian Caffo. 2000. "Simple and Effective Confidence Intervals for Proportions and Differences of Proportions Result from Adding Two Successes and Two Failures." *American Statistician* 54:280–88.

Ajzen, Icek. 1991. "The Theory of Planned Behavior." *Organizational Behavior and Human Decision Processes* 50:179–211.

———. 2001. "The Nature and Operation of Attitudes." *Annual Review of Psychology* 52:27–58.

Baruch, Yehuda. 1999. "Response Rate in Academic Studies—A Comparative Analysis." *Human Relations* 52:421–38.

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* 94:991–1013.

Bobo, Lawrence, Devon Johnson, and Susan A. Suh. 2000. "Racial Attitudes and Power in the Workplace: Do the Haves Differ from the Have-Nots?" Pp. 491–522 in *Prismatic Metropolis: Inequality in Los Angeles*, edited by Lawrence Bobo, Melvin Oliver, James Johnson Jr., and Abel Valenzuela Jr. New York: Russell Sage Foundation.

Bobo, Lawrence and James Kluegel. 1997. "Status, Ideology, and Dimensions of Whites' Racial Beliefs and Attitudes: Progress and Stagnation." Pp. 93–120 in *Racial Attitudes in the 1990s: Continuity and Change*, edited by Steven A. Tuch and Jack K. Martin. Westport, CT: Praeger.

Bobo, L., J. R. Kluegel, and R. A. Smith. 1997. "Laissez Faire Racism: The Crystallization of a 'Kinder, Gentler' Anti-Black Ideology." Pp. 15–42 in *Racial Attitudes in the 1990s: Continuity and Change*, edited by S. Tuch and J. Martin. Westport, CT: Praeger.

Bonczar, Thomas P. 2003. "Prevalence of Imprisonment in the U.S. Population, 1974–2001." *Bureau of Justice Statistics Special Report*. Washingon, DC: U.S. Department of Justice, NCJ 197976.

Bradburn, N. M. 1983. "Response Effects." Pp. 289–328 in *Handbook of Survey Research*, edited by P. Rossi, J. Wright, and A. Anderson. New York: Academic Press.

Bradburn, Norman M., Lance J. Rips, and Steven K. Shevell. 1987. "Answering Autobiographical Questions: The Impact of Memory and Inference on Surveys." *Science* 236:157–61.

Brannon, Robert, Gary Cyphers, Sharlene Hesse, Susan Hesselbart, Roberta Keane, Howard Schuman, Tomas Viccaro, and Diana Wright. 1973. "Attitude and Action: A Field Experiment Jointed to a General Population Survey." *American Sociological Review* 38:625–36.

Bureau of Labor Statistics. 2002. Local Area Unemployment Statistics. (Online Public Data Query). Retrieved March 1, 2003 (http://www.blr.gov/lan/home.htm).

Couch, Kenneth and Mary C. Daly. 2002. "Black–White Wage Inequality in the 1990s: A Decade of Progress." *Economic Inquiry* 40:31–41.

Cox, Brenda G., David A. Binder, B. Nanjamma Chinnappa, Anders Christianson, Michael J. Colledge, and Phillip S. Kott. 1995. *Business Survey Methods*. New York: John Wiley & Sons.

Crandall, C. S. 1994. "Prejudice against Fat People: Ideology and Self-Interest." *Journal of Personality and Social Psychology* 66:882–94.

Crosby, Faye, Stephanie Bromley, and Leonard Saxe.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

1980. "Recent Unobtrusive Studies of Black and White Discrimination and Prejudice: A Literature Review." *Psychological Bulletin* 87:546–63.

Curtin, Richard, Stanley Presser, and Eleanor Singer. 2000. "The Effects of Response Rate Changes on the Index of Consumer Sentiment." *Public Opinion Quarterly* 64:413–28.

Darley, J. M. and P. H. Gross. 1983. "A Hypothesis-Confirming Bias in Labeling Effects." *Journal of Personality and Social Psychology* 44:20–33.

Day, Dianne, Boon Gan, Philip Gendall, and Don Esslemont. 1991. "Predicting Purchase Behaviour." *Marketing Bulletin* 2:18–30.

Deutscher, Irwin. 1966. "Words and Deeds: Social Science and Social Policy." *Social Problems* 13:235–54.

Devine, Patricia. 1989. "Stereotypes and Prejudice: Their Automatic and Controlled Components." *Journal of Personality and Social Psychology* 56:5–18.

Devine, P. G. and A. J. Elliot. 1995. "Are Racial Stereotypes Really Fading? The Princeton Trilogy Revisited." *Personality and Social Psychology Bulletin* 21: 1139–50.

Dovidio, John, Kerry Kawakami, and Samuel Gaertner. 2002. "Implicit and Explicit Prejudice and Interracial Interaction." *Journal of Personality and Social Psychology* 82:62–68.

Downing, David. 1984. "Employer Bias toward the Hiring and Placement of Male Ex-Offenders." Ph.D. dissertation, Department of Education, Southern Illinois University, Carbondale, IL.

Emerson, Michael, George Yancey, and Karen Chai. 2001. "Does Race Matter in Residential Segregation? Exploring the Preferences of White Americans." *American Sociological Review* 66:922–35.

Farley, Reynolds. 1997. "Racial Trends and Differences in the United States 30 Years After the Civil Rights Decade." *Social Science Research* 26:235–62.

Farley, Reynolds, Suzanne Bianchi, and Diane Colasanto. 1979. "Barriers to the Racial Integration of Neighborhoods: The Detroit Case." *Annals of the American Association of Political and Social Sciences* 441:97–113.

Farley, Reynolds, Howard Schuman, Suzanne Bianchi, Diane Colasanto, and Shirley Hatchett. 1978. "Chocolate City, Vanilla Suburbs: Will the Trend toward Racially Separate Communities Continue?" *Social Science Research* 7:319–44.

Farley, Reynolds, Charlotte Steeh, Maria Krysan, Tara Jackson, and Keith Reeves. 1994. "Stereotypes and Segregation: Neighborhoods in the Detroit Area." *American Journal of Sociology* 100:750–80.

Feagin, J. and M. Sykes. 1994. *Living with Racism, the Black Middle Class Experience.* Boston, MA: Beacon Press.

Fishbein, M. 1967. "Attitude and the Prediction of Behavior." Pp. 477–492 in *Readings in Attitude Theory and Measurement*, edited by M. Fishbein. New York: Wiley.

Fishbein, M. and I. Ajzen. 1975. *Belief, Attitude, Intention, and Behavior.* Reading, MA: Addison-Wesley.

Fiske, Susan. 2004. "Predicting Discrimination: A Meta-Analysis of the Racial Attitude-Behavior Literature." Working Paper. Department of Psychology, Princeton University.

Freeman, Richard B., and William M. Rodgers III. 1999. "Area Economic Conditions and the Labor Market Outcomes of Young Men in the 1990s Expansion." Working Paper No. 7073, Cambridge, MA: NBER.

Gans, Herbert. 1992. "Second-Generation Decline: Scenarios for the Economic and Ethnic Futures of the Post-1965 American Immigrants." *Ethnic and Racial Studies* 15: 173–192.

Groves, Robert M. and Lars E. Lyberg. 1988. "An Overview of Nonresponse Issues in Telephone Surveys." Pp. 191–212 in *Telephone Survey Methodology*, edited by Robert M. Groves, Paul P. Biemer, Lars E. Lyberg, James T. Massey, William L. Nicholls II, and Joseph Waksberg. New York: Wiley.

Harrison, R. and C. Bennett. 1995. "Racial and Ethnic Diversity." Pp. 141–210 in *State of the Union, America in the 1990s,* vol. 2, *Social Trends 1990 Monograph Series*, edited by R. Farley. New York: Russell Sage Foundation.

Heckman, James. 1998. "Detecting Discrimination." *The Journal of Economic Perspectives* 12: 101–116.

Heckman, James and Peter Seligman. 1993. "The Urban Institute Audit Studies: Their Methods and Findings." Pp. 187–258 in *Clear and Convincing Evidence: Measurement of Discrimination in America*, edited by Michael Fix and Raymond J. Struyk. Washington, DC: Urban Institute Press.

Holzer, Harry. 1996. *What Employers Want: Job Prospects for Less-Educated Workers.* New York: Russell Sage Foundation.

Holzer, Harry and David Neumark. 2000. "What Does Affirmative Action Do?" *Industrial and Labor Relations Review* 53: 240–71.

Holzer, Harry, Steven Raphael, and Michael Stoll. 2003. "Employer Demand for Ex-Offenders: Recent Evidence from Los Angeles." Discussion Paper No. 1268–03, Institute for Research on Poverty, University of Wisconsin, Madison, WI.

Holzer, Harry, Paul Offner and Elaine Sorensen. 2005. "What Explains the Continuing Decline in Labor Force Activity among Young Black Men?" *Labor History* 46: 37–55.

Holzer, Harry and Michael Stoll. 2001. *Employers and Welfare Recipients: The Effects of Welfare*

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

From callbacks to labor markets - In-person audit, New York City (Pager and Western 2009)

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Reform in the Workplace. San Francisco, CA: Public Policy Institute of California.

Husley, Lonnie Freeman. 1990. "Attitudes of Employers with Respect to Hiring Released Prisoners." Ph.D. dissertation, Department of Industrial Education, Texas A&M University, TX.

Jensen, W. and W. C. Giegold. 1976. "Finding Jobs for Ex-Offenders: A Study of Employers' Attitudes." *American Business Law Journal* 14:195–225.

Kinder, D. and D. O. Sears. 1981. "Prejudice and Politics: Symbolic Racism Versus Racial Threats to the Good Life." *Journal of Personality and Social Psychology* 40:414–31.

Kirschenman, Joleen and Katherine Neckerman. 1991. "'We'd Love to Hire Them, but . . .': The Meaning of Race for Employers." Pp. 203–232 in *The Urban Underclass*, edited by Christopher Jencks and P. Peterson. Washington, DC: Brookings Institute.

Krysan, Maria. 1999. "Prejudice, Politics, and Public Opinion: Understanding the Sources of Racial Policy Attitudes." *Annual Review of Sociology* 26:135–68.

———. 2002. "Whites Who Say They'd Flee: Who Are They, and Why Would They Leave?" *Demography* 39:675–96.

Kutner, Bernard, Carol Wilkins, and Penny Rechtman Yarrow. 1952. "Verbal Attitudes and Overt Behavior Involving Racial Prejudice." *Journal of Abnormal Social Psychology* 47:649–52.

LaPiere, Richard T. 1934. "Attitudes vs Actions." *Social Forces* 13:230–37.

Lichtenstein, Sarah and J. Robert Newman. 1967. "Empirical Scaling of Common Verbal Phrases Associated with Numerical Probabilities." *Psychonomic Science* 9:563–64.

Mare, R. D. 1995. "Changes in Educational Attainment and School Enrollment." Pp.155–213 in *State of the Union: America in the 1990s,* vol. I, *Economic Trends*, edited by R. Farley. New York: Russell Sage Foundation.

Massey, Douglas and Garvey Lundy. 2001. "Use of Black English and Racial Discrimination in Urban Housing Markets: New Methods and Findings." *Urban Affairs Review* 36:452–69.

Merton, Robert K. 1949. "Discrimination and the American Creed." Pp. 99–126 in *Discrimination and the National Welfare*, edited by Robert M. MacIver. New York: Harper and Row.

Moss, Philip and Chris Tilly. 2003. *Stories Employers Tell: Race, Skill, and Hiring in America*. New York: Russell Sage Foundation.

National Research Council. 2004. *Measuring Racial Discrimination.* Panel on Methods for Assessing Discrimination, Committee on National Statistics, Division of Behavior and Social Sciences and Education, edited by Rebecca M. Blank, Marilyn Dabady, and Constance F. Citro. Washington, DC: The National Academies Press.

Pace, C. Robert and Jack Friedlander. 1982. "The Meaning of Response Categories: How Often Is 'Occasionally,' 'Often,' and 'Very Often'"? *Research in Higher Education* 17:267–81.

Pager, Devah. 2003. "The Mark of a Criminal Record." *American Journal of Sociology* 108:937–75.

———. 2002. "The Mark of a Criminal Record." Doctoral dissertation. Department of Sociology, University of Wisconsin-Madison.

Pettit, Becky and Bruce Western. 2004. "Mass Imprisonment and the Life Course: Race and Class Inequality in U.S. Incarceration." *American Sociological Review* 69:151–69.

Poskocil, A. 1977. "Encounters Between Blacks and White Liberals: The Collision of Stereotypes." *Social Forces* 55:715–27.

Remington, Todd D. 1992. "Telemarketing and Declining Survey Response Rates." *Journal of Advertising Research* 32:6–7.

Saenger, Gerhart and Emily Gilbert. 1950. "Custom Reactions to the Integration of Negro Sales Personnel." *International Journal of Opinion and Attitude Research* 4:57–76.

Sagar, H. and J. W. Schofield. 1980. "Racial and Behavioral Cues in Black and White Children's Perception fo Ambiguously Aggressive Acts." *Journal of Personality and Social Psychology* 39:590–98.

Schaeffer, Nora Cate. 1991. "Hardly Ever or Constantly? Group Comparisons Using Vague Quantifiers." *Public Opinion Quarterly* 55:395–423.

Schuman, Howard. 1995. "Attitudes." Pp. 68–89 in *Sociological Perspectives on Social Psychology*, edited by Karen Cook, Gary Fine, and James House. Boston, MA: Allyn & Bacon.

Schuman, Howard and Lawrence Bobo. 1988. "Survey-Based Experiments on White Racial Attitudes toward Residential Integration." *American Journal of Sociology* 94:273–99.

Schuman, Howard and Michael P. Johnson. 1976. "Attitudes and Behaviors." *Annual Review of Sociology* 2:161–207.

Schuman, Howard, Charlottee Steeh, Lawrence Bobo, and Maria Krysan. 2001. *Racial Attitudes in America: Trends and Interpretations*. Rev. ed. Cambridge, MA: Harvard University Press.

Schutz, Heidi and Bernd Six. 1996. "How strong is the Relationship Between Prejudice and Discrimination? A Meta-Analytic Answer." *International Journal of Intercultural Relations* 20: 441–62.

Smith, Tom W. 1991. *Ethnic Images: General Social Survey Technical Report, 19*. Chicago, IL: National Opinion Research Center, University of Chicago.

Sniderman, Paul M. and Douglas B. Grob. 1996.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

"Innovation in Experimental Design in Attitude Surveys." *Annual Review of Sociology* 22:377–99.

Sniderman, Paul M. and Thomas Piazza. 1993. *The Scar of Race*. Cambridge, MA: The Belknap Press of Harvard University Press.

Traugott, Michael W. and John P. Katosh. 1979. "Response Validity in Surveys of Voting Behavior." *Public Opinion Quarterly* 43:359–77.

Turner, Margery, Michael Fix, and Raymond Struyk. 1991. *Opportunities Denied, Opportunities Diminished: Racial Discrimination in Hiring*. Washington, DC: Urban Institute Press.

Waldinger, Roger and Michael Lichter. 2003. *How the Other Half Works: Immigration and the Social Organization of Labor*. Berkeley, CA: The University of California Press.

Western, Bruce and Becky Pettit. Forthcoming. "Black-White Wage Inequality, Employment Rates, and Incarceration." *American Journal of Sociology*.

Wicker, A. W. 1969. "Attitudes versus Actions: The Relationship of Verbal and Overt Behavioral Responses to Attitude Objects." *Journal of Social Issues* 25:41–78.

Wilson, William Julius. 1978. *The Declining Significance of Race*. Chicago, IL: University of Chicago Press.

———. 1996. *When Work Disappears: The World of the New Urban Poor*. New York: Vintage Books.

Word, C. O., M. P. Zanna, and J. Cooper. 1974. "The Nonverbal Mediation of Self-Fulfilling Prophecies in Interracial Interactions." *Journal of Experimental Social Psychology* 10:109–20.

Yinger, John. 1995. *Closed Doors, Opportunities Lost*. New York: Russell Sage Foundation.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

From callbacks to labor markets - in-person audit, New York City (Pager and Western 2009)

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.