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.

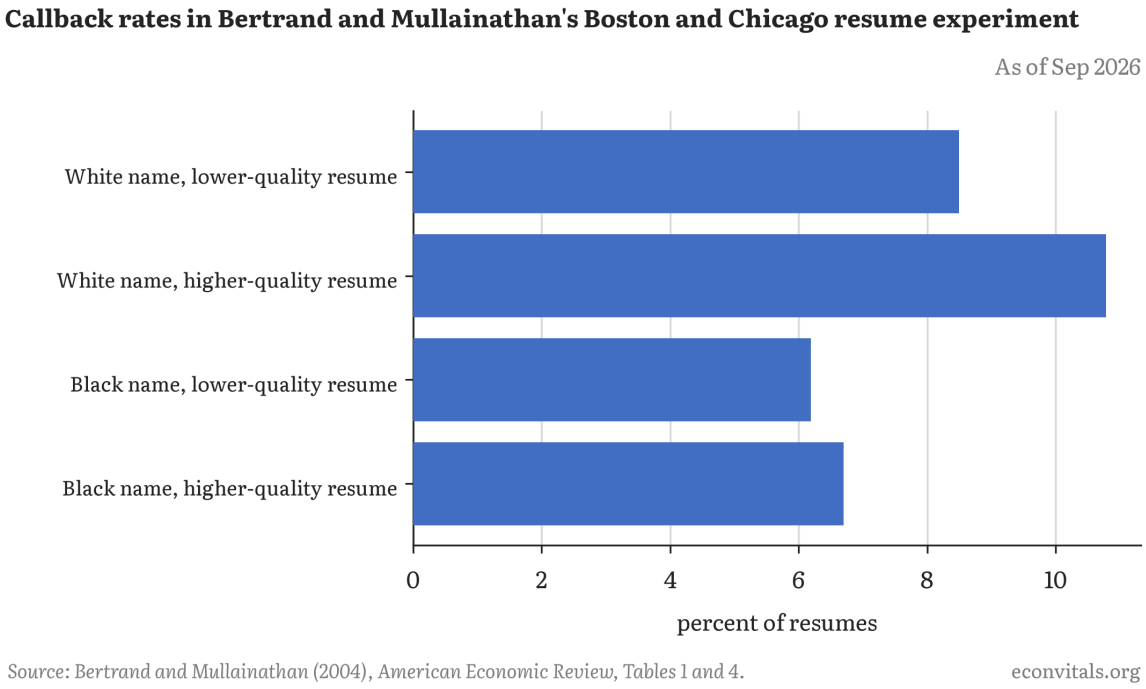


Figure 1. White-named resumes drew a callback 9.65 percent of resumes of the time against 6.45 percent of resumes for otherwise identical Black-named resumes. Upgrading the resume raised white callbacks by 2.29 percentage points but Black callbacks by only 0.51 percentage points.

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.

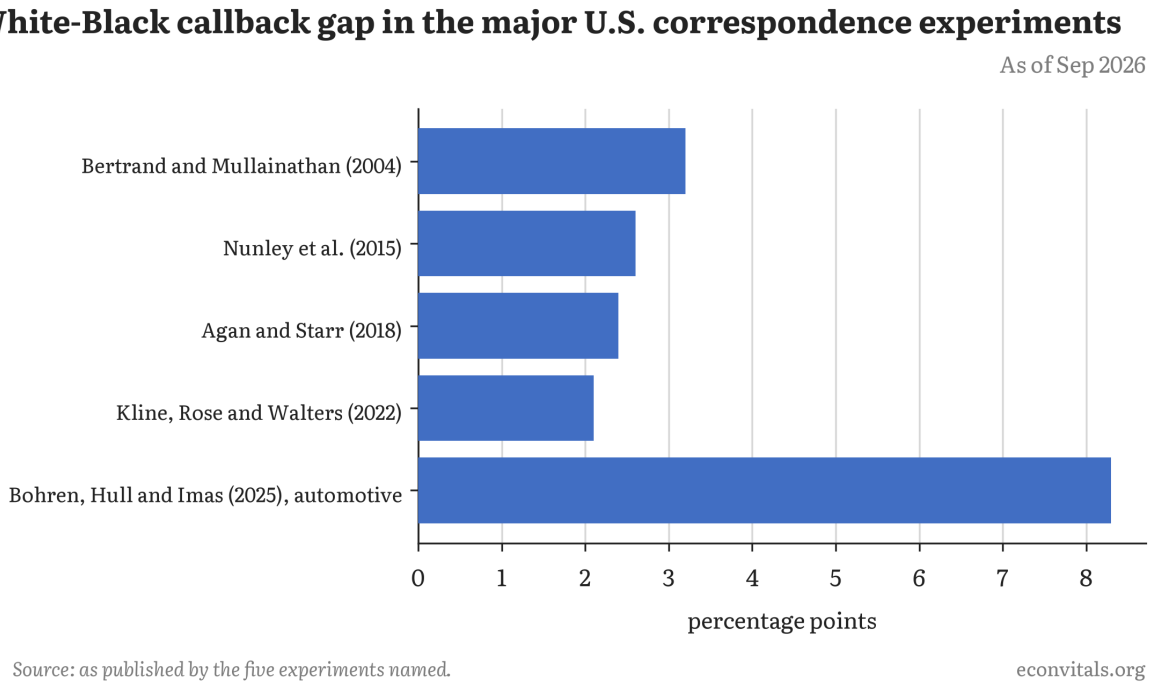


Figure 2. The level gap replicates across two decades: 3.20 percentage points in the original experiment, 2.1 percentage points in the largest one. The proportional gap moves with the base callback rate, which in Kline, Rose and Walters was 3 times the original study's.

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.

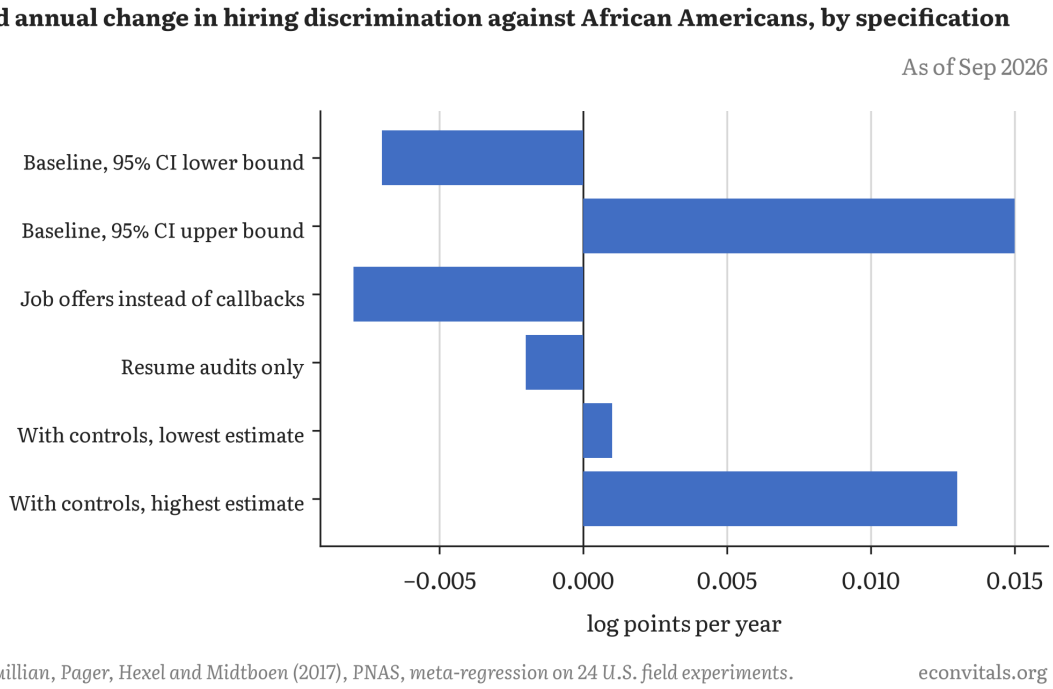


Figure 3. Across every specification the meta-regression can neither reject a flat trend nor a rising one; even the most favorable bound implies a decline of only 0.7 percent per year in the white-Black callback ratio.
Source: Quillian, Pager, Hexel and Midtboen (2017), PNAS, meta-regression on 24 U.S.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

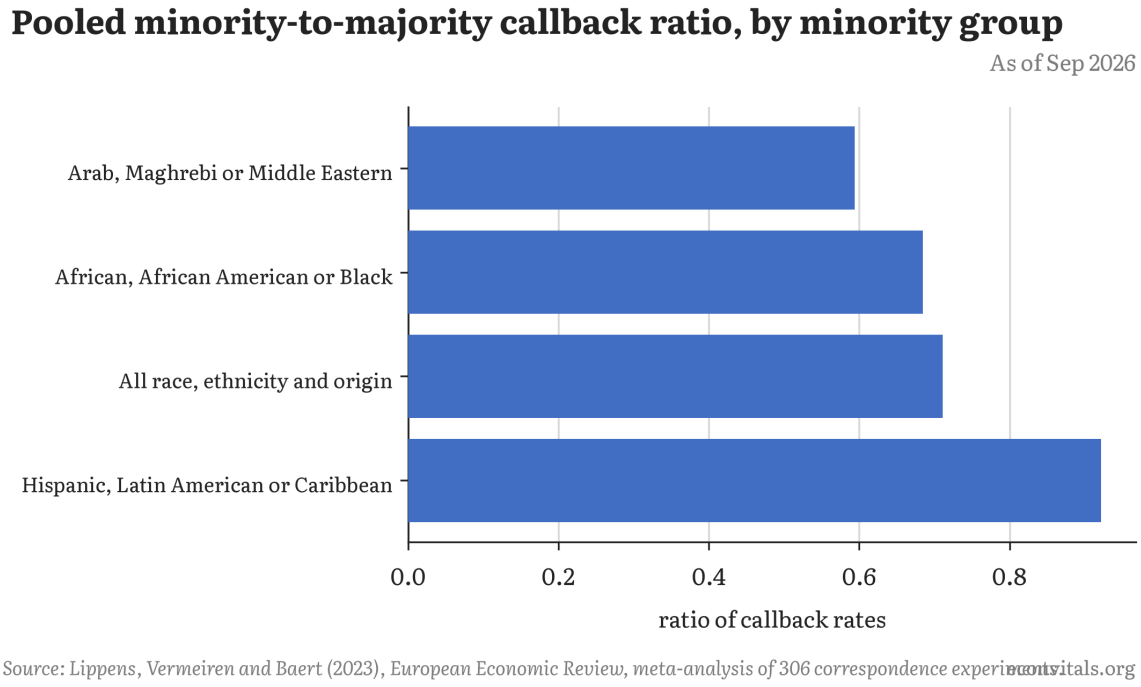


Figure 4. The register's pooled ethnic ratio is 0.7113 ratio, but the average conceals its parts: Arab and Maghrebi applicants face a ratio of 0.5937 ratio while Hispanic applicants show no significant penalty at all (0.9220 ratio).

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.

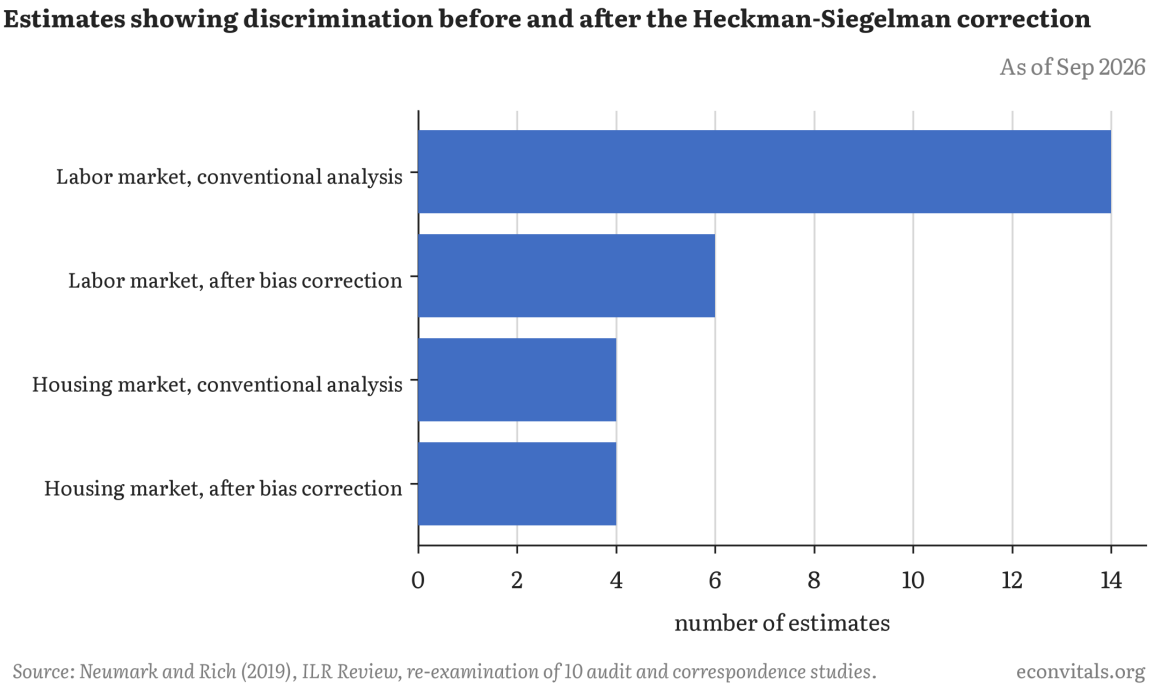


Figure 5. All 14 estimates labor-market estimates show discrimination when analyzed conventionally; 6 of 14 estimates survive the correction. In the housing studies every corrected estimate still shows significant discrimination.

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.

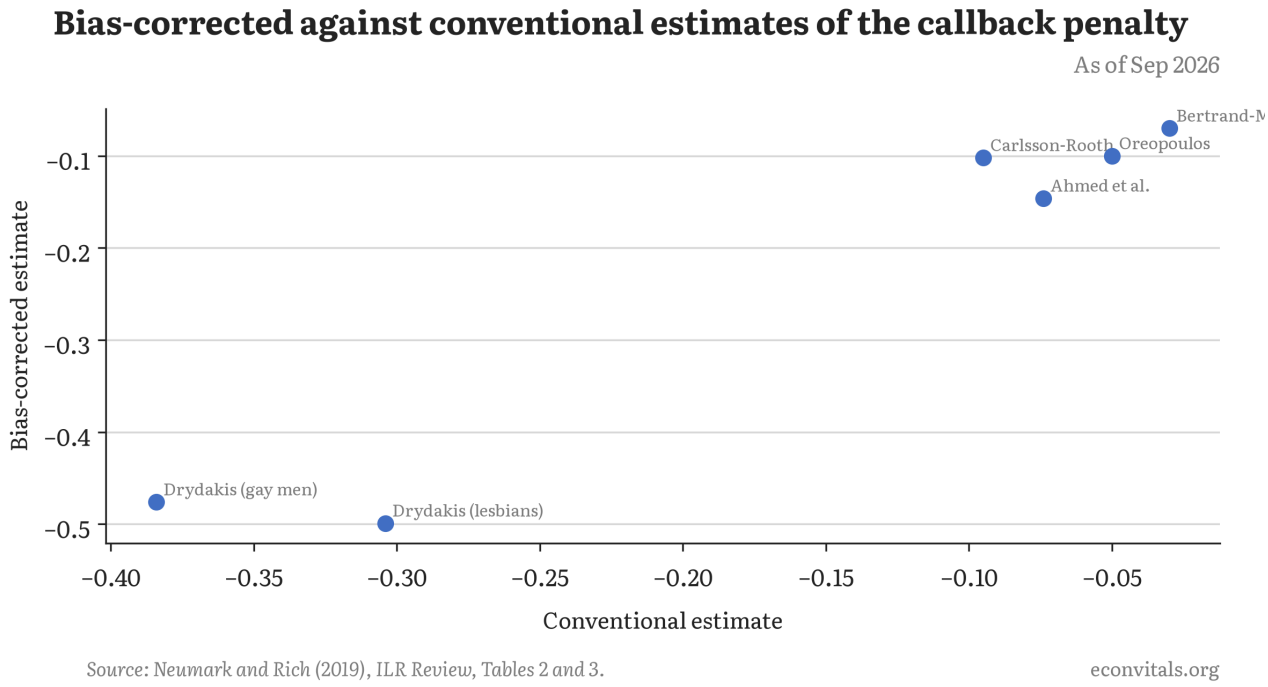


Figure 6. Where the correction can be computed it usually makes discrimination look worse, not better: Bertrand and Mullainathan's penalty moves to -0.070 probability of a call-back and Oreopoulos's to -0.10 probability of a call-back. The two reversals are elsewhere - Baert et al. flips sign and loses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

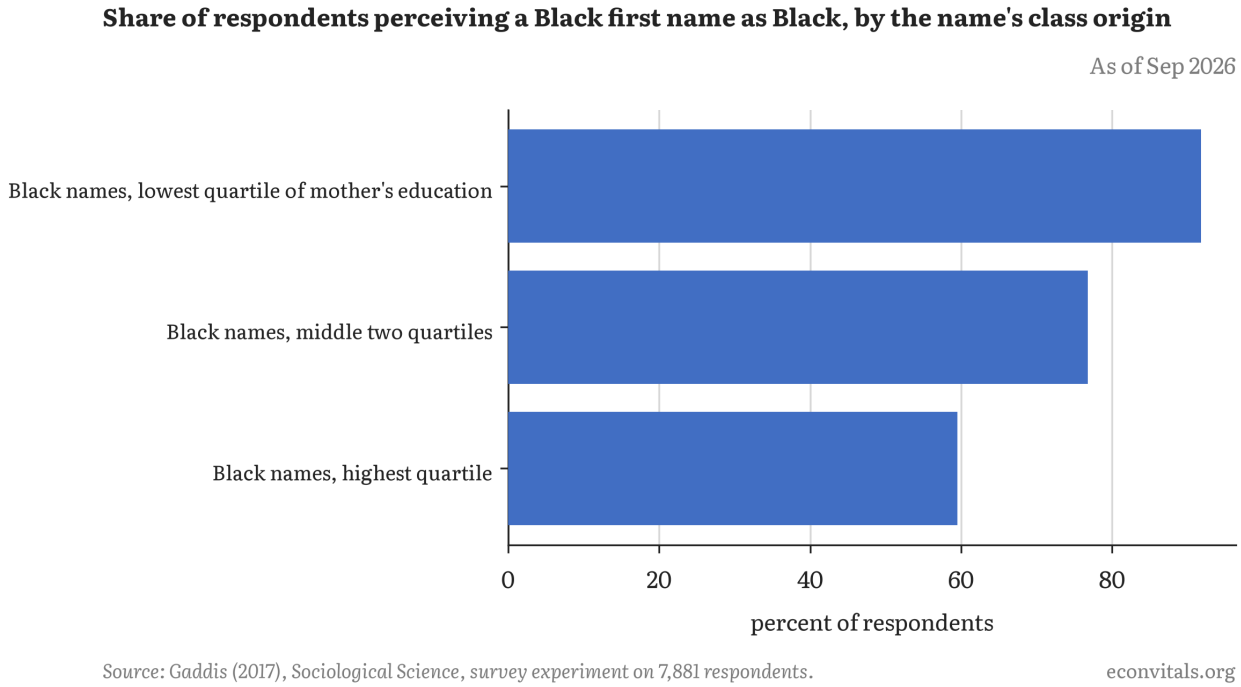


Figure 7. Names given disproportionately by the least educated mothers are read as Black by 91.8 percent of respondents; names from the most educated quartile by only 59.5 percent. The names that carry the cleanest racial signal are the lowest-status ones.

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.

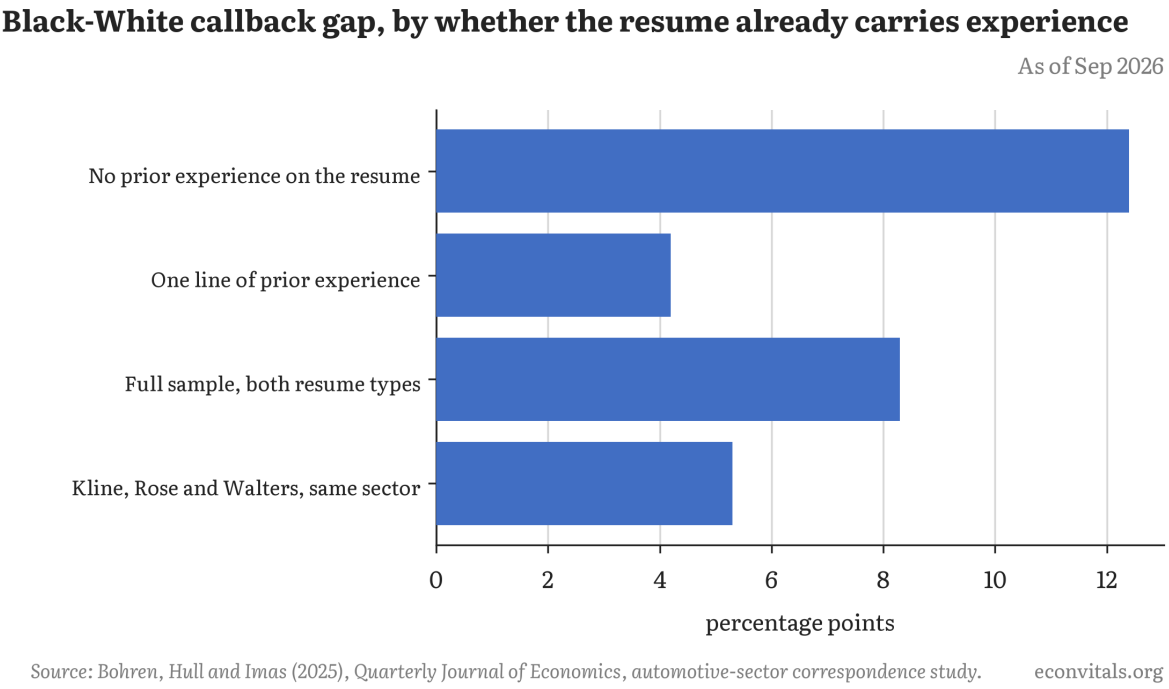


Figure 8. Holding the resume fixed at no experience the gap is 12.4 percentage points; adding one line of experience shrinks it to 4.2 percentage points, which is not significant. Because a conventional audit conditions on experience, reading only its estimate understates total discrimination by about 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).

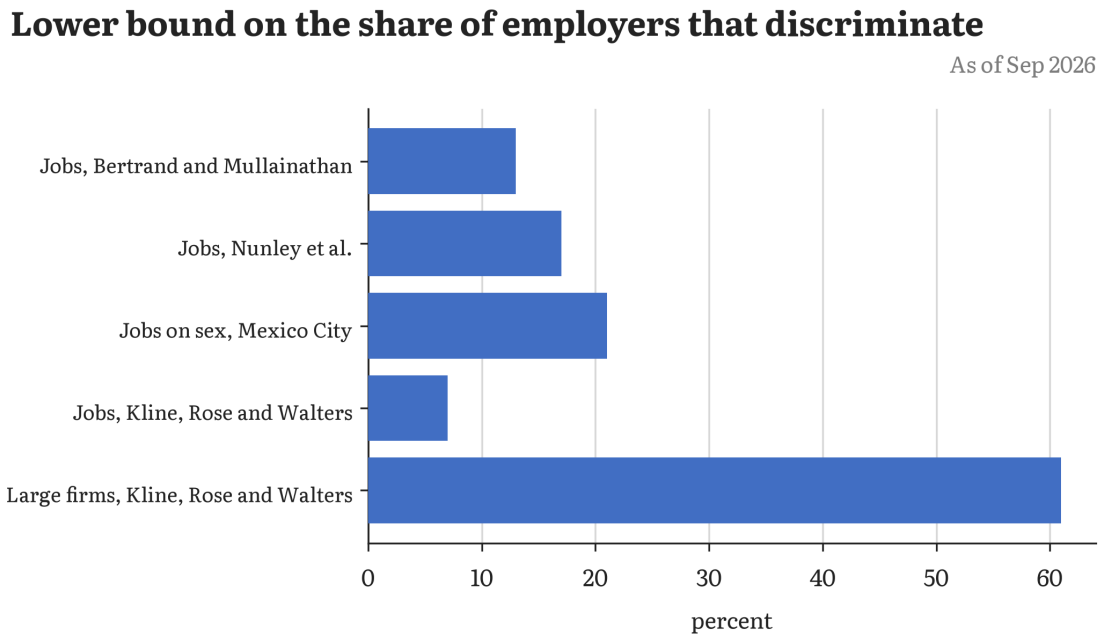


Figure 9. The average gap is produced by a minority of workplaces: at least 13 percent of jobs in the original experiment and 17 percent of jobs in Nunley et al., yet at least 61 percent of the 108 large firms in the biggest experiment discriminate somewhere in their hiring.

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.

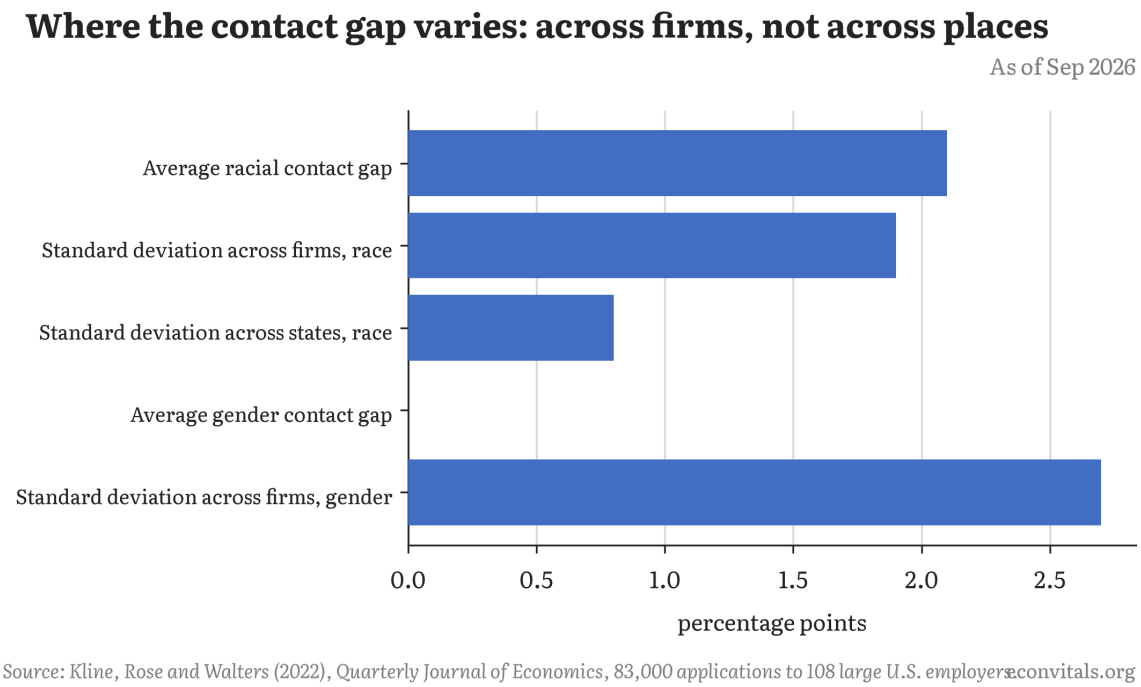


Figure 10. The between-firm standard deviation of racial contact gaps, 1.9 percentage points, is almost as large as the average gap itself, while the spread across states is 0.8 percentage points. The average gender gap is essentially zero and its between-firm spread is 2.7 percentage points.

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.

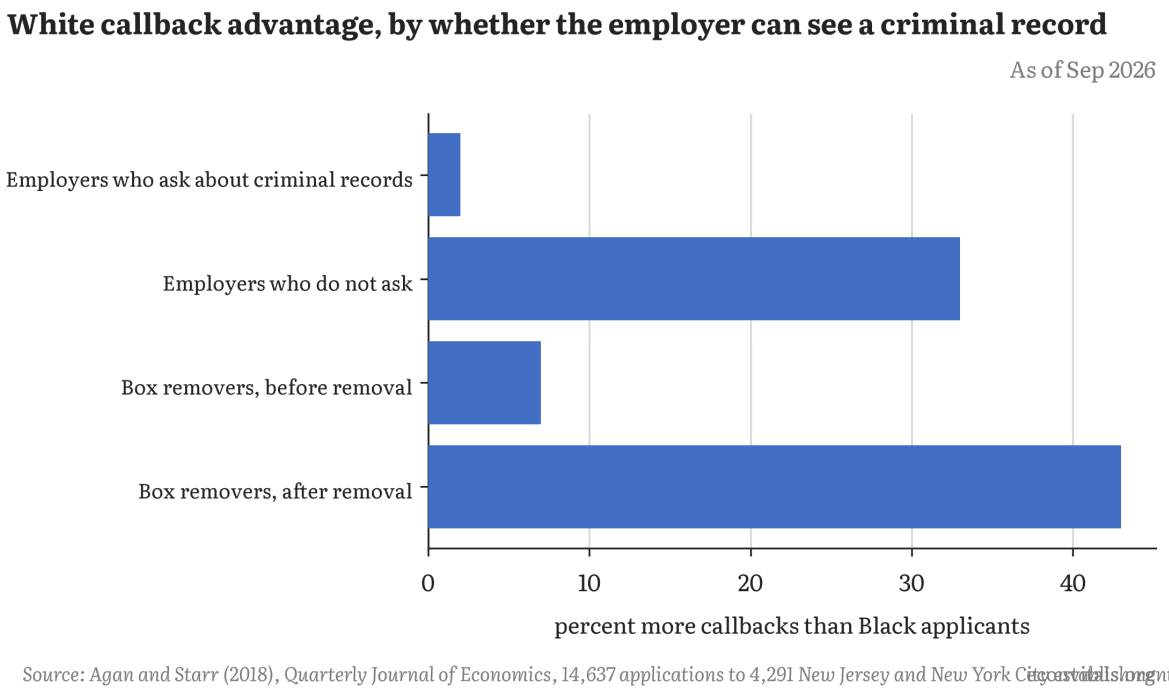


Figure 11. Where employers ask about records the race gap is 0.3 percentage points; where they do not, white applicants receive 33 percent. The same employers before and after dropping the question show the gap widening by 3.6 percentage points.

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