Evidence from Field Experiments in Hiring Shows Substantial Additional Racial Discrimination after the Callback

Lincoln Quillian, John J. Lee and Mariana Oliver, *Northwestern University*

Field experiments using fictitious applications have become an increasingly important method for assessing hiring discrimination. Most field experiments of hiring, however, only observe whether the applicant receives an invitation to interview, called the "callback." How adequate is our understanding of discrimination in the hiring process based on an assessment of discrimination in callbacks, when the ultimate subject of interest is discrimination in job offers? To address this question, we examine evidence from all available field experimental studies of racial or ethnic discrimination in hiring that go to the job offer outcome. Our sample includes 12 studies encompassing more than 13,000 job applications. We find considerable additional discrimination in hiring after the callback: majority applicants in our sample receive 53% more callbacks than comparable minority applicants, but majority applicants receive 145% more job offers than comparable minority applicants. The additional discrimination from interview to job offer is weakly correlated ($r = 0.21$) with the level of discrimination earlier in the hiring process. We discuss the implications of our results for theories of discrimination, including statistical discrimination.

Introduction

A major advance in the social science literature on race and ethnicity has been the development of field experimental methods to measure discrimination. Field experiments allow investigators to combine the high internal (causal) validity of experiments with the high external validity of being conducted "in the field"

Address correspondence to: Lincoln Quillian, Department of Sociology, Northwestern University, 1810 Chicago Avenue, Evanston IL 60202. E-mail: l-quillian@northwestern.edu. *A previous version of this paper was presented at the April 2018 meetings of the Southern Sociological Society in New Orleans, LA. We received helpful comments from Michael Gaddis, Arnfinn Midtbøen, and the Social Forces reviewers. We gratefully acknowledge financial support for this project from the Russell Sage Foundation, Award #88-15-06. Data and code for this study is available at https://sites.northwestern.edu/dmap*

Social Forces 99(2) 732–759, December 2020
doi:10.1093/sf/soaa026
Advance Access publication on 7 April 2020

rather than in a laboratory (National Research Council 2004; Gerber and Green 2012). Unlike reports of discrimination from surveys, field experiments are grounded in actual behavior, avoiding problems of weak attitude–behavior correspondence (Pager and Quillian 2005) and social desirability bias (Gaddis 2018). As field experiments have grown in popularity, a large literature has developed: there are now more than one hundred field experimental studies of hiring discrimination against minority races and ethnicities across more than twenty countries.[1]

Despite the widespread use of field experiments to study hiring discrimination, the vast majority of field experimental studies do not observe whether a job offer is extended to candidates.[2] Instead, most studies stop with whether or not an applicant receives an invitation to interview, often referred to as a "callback." Although researchers are interested in explaining racial disparities in hiring outcomes, the callback is used as a proxy for the job offer because it is much more difficult to conduct a field experiment that goes all the way to the job offer outcome. Field experiments that go to the job offer require extensive time for training auditors and for each applicant to go through the entire hiring process, and run into ethical concerns because they use significant amounts of employer time without their consent (Cherry and Bendick 2018).

Given these difficulties, the widespread use of callbacks in audit studies is understandable. Nevertheless, it leaves unanswered some important questions: Does the focus on callbacks miss significant additional discrimination later in the hiring process—discrimination from interview to job offer? If so, how does this "extra" discrimination affect our understanding of the nature and causes of racial and ethnic discrimination in hiring?

Although specifics vary, a job search often proceeds through two stages. In the first stage, applicants submit an application or resume; if successful, they receive a callback for an interview. In the second stage, applicants who received a callback are interviewed and, if successful, receive a job offer. Discrimination may be present in the second stage for several reasons: employers may miss clues indicating an applicant's race in the resume; employers may respond differently to race in the face-to-face context of the interview than in the abstracted context of reading a resume; the decision maker from interview to job offer may be different than the person selecting candidates to interview; or employers may select employees in the interview stage in part based on "cultural fit" in ways that produce discrimination.

Racial discrimination in hiring contributes to employment gaps between majority and minority populations, impedes the social and economic incorporation of immigrants and has negative psychological and health effects on the targets of discrimination (Williams and Jackson 2005; Attström 2007; Pascoe and Richman 2009). Given the high costs of racial discrimination in hiring, it is important to understand the full picture of how discrimination operates.

We assess the adequacy of callbacks for understanding discrimination in job offers by comparing callback and job offer outcomes in all field experimental studies of racial discrimination in hiring that go to the job offer. To make this comparison we use techniques from the meta-analysis literature, the branch

of statistics concerned with combining results across studies (Borenstein et al. 2009). Our results indicate that substantial, additional racial discrimination occurs even after minority candidates make it to the interview stage. Because of this, studies that only use callbacks seriously underestimate the complete extent of discrimination in the hiring process.

Background

Overview of Field Experiments of Discrimination in Hiring

In field experiments of racial discrimination in hiring, fictitious applicants from different racial or ethnic groups apply for jobs. By managing the applicant profiles to match applicants on background characteristics, investigators can largely rule out differences other than employer discrimination as explanations of racial differences in the hiring outcome. Because they can identify discrimination confidently, field experiments have become a key social science tool for understanding the extent of discrimination in hiring. Some field experiments have been performed in-person, which we call in-person audits, whereas others have been conducted through the mail or over the internet, which we call resume audits.

For in-person audits, researchers send teams of trained actors to apply for the same job vacancies (e.g., Attström 2007; Pager et al. 2009). Each team includes at least one actor belonging to the native or dominant racial group and another from a racial minority group. Teams are assigned equivalent fictitious employment credentials like education, training and previous experience. The majority and minority actors undergo a period of training that involves practice calls to employers, mock interviews and standardizing candidate responses to interview questions (Bendick et al. 2010). Actors are matched based on physical appearance, age and demeanor. In-person audit studies usually rely on at least two signals about the applicant's race: the applicant's name in the resume and the applicant's in-person and physical appearance.[3]

For resume audit studies, researchers submit resumes representing fictitious applicants by mail or over the internet (e.g., Gaddis 2015). Applicants from the majority and minority groups are given resumes with on-average equivalent qualifications; some audits randomly assign attributes to ensure that there are no systematic differences between majority and minority groups. The applicant's race or ethnic background is usually signaled by the applicant's name on the resume (Agerström et al. 2012; Bursell 2014). For instance, in U.S. studies, researchers have used white-sounding names such as "Emily Walsh" or "Greg Baker" to signal the race of white applicants, and distinctively African-American names such as "Lakisha Washington" or "Jamal Jones" to send signals about the race of black applicants (Bertrand and Mullainathan 2004; on name signaling see Gaddis 2017a; Butler and Homola 2017).

In all resume audit and many in-person audit studies, the main outcome of interest is the callback for an interview (Baert 2018; Zschirnt and Ruedin 2016). If an applicant receives a callback, the outcome is recorded and the

auditor either declines the invitation to interview or does not respond. A callback received signals indication of employer interest and hence of success in the hiring process.

However, a limited number of in-person audit studies have pursued applications all the way to the job offer outcome.[4] We use these studies to evaluate how callback and job offer outcomes correspond. The level of discrimination at the callback stage is only a perfect proxy for the total discrimination in hiring if there is no further discrimination (and no "reverse" discrimination) at the final stage, when employers are deciding whether to extend a job offer.

Discrimination over the Hiring Process

Previous scholarship has identified several explanations for why employers might favor a majority candidate over an equally qualified or even superior minority candidate. A relevant distinction developed in the economics literature is between "taste-based" and "statistical" discrimination (Guryan and Charles 2013; Neumark 2018). In its basic form, taste-based discrimination is grounded in the preferences of employers for and against employees of certain races, whereas statistical discrimination entails employers making decisions grounded in race-based, statistical evaluations of potential employees.

Although these are useful distinctions, from a sociological standpoint they are underspecified in capturing the various factors that drive discrimination. The preferences underpinning taste-based discrimination can encompass many specific forms of racism and prejudice. For instance, employers may hold prejudices against racial and ethnic minorities rooted in suspicions of or hostility toward foreign cultural norms, values or attitudes (Pager and Shepherd 2008). On the other hand, biases that affect hiring may be unconscious, as demonstrated by studies of "implicit" racial attitudes (Greenwald et al. 1998; Rooth 2010).

In statistical discrimination, employers use the average characteristics of employees from different racial groups to draw conclusions about individual prospective employees based on their race (Arrow 1973). One implication of statistical discrimination is that the level of discrimination should respond to information about applicants: as more individual information about applicants becomes available, employers should rely less heavily on race-based statistical profiles, leading to reduced levels of discrimination (Altonji and Pierret 2001). Although some sociological theories also view statistical discrimination as responding in part to incomplete information about applicants, they argue that these beliefs can be on-average incorrect, because of the negative cultural stereotyping of minority group members that employers might engage in (e.g., Bobo et al. 2012; Quillian and Pager 2010; Kirschenman and Neckerman 1991).

Theories of why employers discriminate provide little clear guidance on how much discrimination exists at *each stage* of the hiring process. However, as we elaborate below, there are reasons to believe that the level of discrimination may differ across stages. Accordingly, we assess levels of discrimination at each stage of hiring: (1) application to callback and (2) interview to job offer.

Perhaps the most compelling reason to expect more discrimination from application to callback than from interview to job offer is due to selective attrition across the hiring process. The callback comes before the interview, and interviews generally only occur for applicants who receive a callback. This suggests that minority applicants who advance to the interview stage are more likely to do so with employers who have a relatively low propensity to discriminate: employers with a high propensity to discriminate are likely to weed out identifiable minority applicants at the earlier callback stage.[5]

A second reason that discrimination might be low from interview to job offer is because the additional information employers receive about candidates during the interview may reduce statistical discrimination. Statistical discrimination theory posits that employers rely on the "average" profiles of groups to compensate for any lack of individual applicant information. The interview provides additional individual information about an applicant's speech, dress, appearance and also an opportunity for the employer to ask about the applicant's background, such as past work history. If statistical discrimination on these attributes is the primary basis for employer discrimination, then this additional information should result in employers relying less on group averages when making hiring decisions, resulting in less discrimination from interview to job offer (Altonji and Pierret 2001).

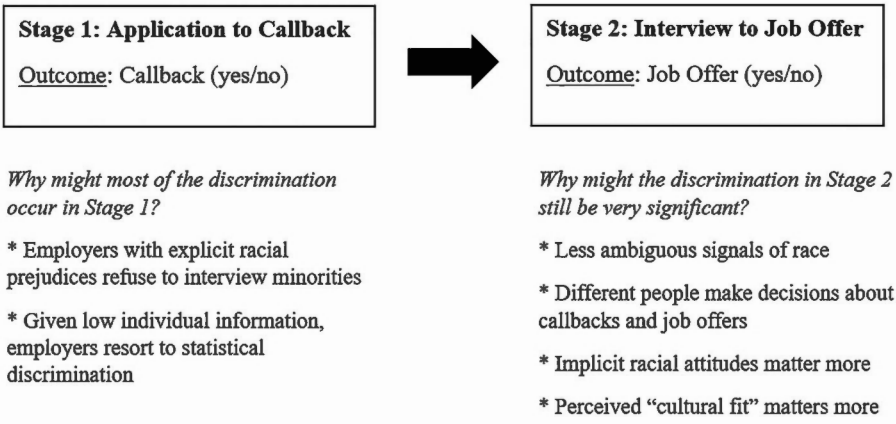
Other theories, however, suggest just the opposite pattern: discrimination at the job offer stage might actually be quite significant. First, there may be substantial discrimination from interview to job offer because race is more obvious in the interview situation. Racially-distinct names provide good but imperfect signals about ethnicity (Gaddis 2017a, 2017b), and the clarity of these signals also depends on contextual factors like geography (Crabtree and Chykina 2018). Employers with racial prejudices obviously cannot discriminate in callbacks if they are unable to recognize the signal of race or ethnicity via the applicant's name. Appearance in interviews, on the other hand, sends a clear signal that is less likely to be misunderstood, and as a result may produce more discrimination.[6]

Second, in some cases the decision makers in the callback and job offer stages are different people. For instance, the human resources division may select interviewees but a company supervisor may conduct the interviews and make decisions about job offers. The existence of two different decision makers will tend to produce a weaker link between stages, and the selection process favoring lower discrimination in the second stage may not operate as expected.

Third, even when employers are able to correctly infer an applicant's race from application materials, racial stereotypes or attitudes may be more strongly invoked by face-to-face interactions than the more abstracted situation of seeing a name on a resume. Many tests for implicit attitudes such as the Implicit Attitudes Test (IAT) use images of individuals from different racial and ethnic groups, which suggests that the general salience of race might be heightened in the context of face-to-face interactions (Greenwald et al. 1998).

Fourth, in-person interaction allows employers to evaluate more subjective cultural attributes and styles of interaction. Rivera (2012) finds that "cultural

Figure 1. Expectations of discrimination by stage.



fit" is a key dimension evaluated by employers in interviews at professional service firms. Nonwhite applicants who tend to have different cultural attributes (e.g., types of dress or presentation styles) but are otherwise equally qualified for the job may thus face additional disadvantages during the interview.[7] This is underscored by a recent study of hiring in Germany by Weichselbaumer (2016), which found much higher discrimination against a Turkish woman who wore a headscarf in photos submitted for jobs compared to the same applicant without a headscarf (in Germany it is typical to submit a photo in a job application). In cases where wearing a headscarf is not relevant to job performance, this can be viewed as a type of ethnic discrimination.

Given the foregoing discussion, it is unclear whether racial discrimination occurs primarily during the initial application stage, or is also substantial during the interview stage.[8] Figure 1 provides a summary of the theoretical reasons for each view. To assess how our view of total discrimination is altered by also considering the job offer, rather than just the callback, we design a meta-analysis that examines all of the field experimental studies of hiring that go to the job offer outcome and compare levels of discrimination at the callback and job offer stages.

Previous Work Contrasting Callback and Job Offer Outcomes

We know of only one previous study that compares callback and job offer outcomes: a monograph by Zegers de Beijl (2000), which discusses results from three in-person audit studies that go to the job offer. Using these three audit studies, Zegers de Beijl (2000) contrasts the prevalence of discrimination at three different stages of hiring: a pre-application inquiry as to whether the job is still available, the callback for an interview and then the job offer. He concludes that the level of discrimination is highest at the initial pre-interview stage and then declines across stages.

An important caveat to Zegers de Beijls (2000) conclusions, however, is that he defines the prevalence of discrimination at a given stage as the number of cases of unequal advancement relative to the total number of applicants who initially applied. This approach confounds the number of persons at risk of being discriminated against at a given hiring stage with the rate of discrimination. As applicants are weeded out across stages of hiring, the number of applicants who could face discrimination shrinks, automatically contributing to declining rates of discrimination as Zegers de Beijl measures it.

For instance, suppose one hundred majority–minority auditor pairs initially apply for a job. At the callback stage, in forty pairs both the majority and minority auditor receive a callback, in twenty pairs only the majority auditor receives a callback, in five pairs only the minority auditor receives a callback, and in the remaining thirty-five pairs neither auditor receive a callback. Zegers de Beijl (2000) estimates discrimination at the callback stage as $\frac{20-5}{100} = 15\%$. At the job offer stage, only the forty pairs of auditors who both receive callbacks go on to the interview. Suppose in ten cases both receive a job offer, in ten cases only the majority auditor receives a job offer, and in the remaining twenty cases neither auditor receives a job offer. Zegers de Beijl (2000) would compute discrimination at the job offer stage as $\frac{10}{100} = 10\%$. Discrimination then appears to have declined over stages from 15% to 10%, but this reflects the fact that only forty pairs of auditors were interviewed (made it to the second stage) and so were at risk of discrimination at that stage.[9]

Zegers de Beijl (2000) assesses discrimination as occurring whenever the majority applicant gets further in the hiring process than the minority applicant. By contrast, we assess discrimination based on the final outcome observed, job offers, and relative to the at-risk pool at each stage. We prefer to focus on discrimination in job offers, rather than on whether a majority candidate makes it further in the process than his minority counterpart, because it is more closely linked to racial disparities in hiring and employment. Using this method, we come to a conclusion quite different to that of Zegers de Beijl: we find fairly similar levels of discrimination from interview to job offer as from application to callback, rather than evidence that most discrimination occurs at the first stage of hiring.

Methods

Our research design followed a three-step process. First, we identified all of the field experimental studies of racial discrimination in hiring that pursue applications to the job offer. Second, we coded the studies using a coding rubric and created a database of results, which included counts of applications, callbacks and job offers by racial/ethnic group. Third, we performed a statistical analysis of the data using meta-analytic methods (Borenstein et al. 2009). Meta-analysis is a set of statistical techniques used to aggregate information across multiple existing studies to produce an overall estimate of an effect of interest.

The search for studies was part of a larger project to gather and code information from all existing field experiments of racial and ethnic discrimination

in hiring. This included a search of major databases for field experiments of hiring, a search of cited references, and an e-mail survey of authors of field experiments asking for field experiments of discrimination, including unpublished studies. Details of these procedures are discussed in Supplementary Material Appendix B. More than one hundred field experimental studies were identified using these methods. However, only thirteen of these studies proceeded all the way to the job offer, and one of these was subsequently excluded for lacking a comparable callback outcome (see Supplementary Material Appendix B), leaving a total base sample of twelve studies.

Table 1 lists the studies in our sample and, in the third column, the minority groups that are the targets of discrimination.[10] Several of the studies include multiple target groups, such as blacks and Latinos in the James and DelCastillo (1992) study. Because of this, the twelve studies include fifteen estimations of discrimination against minority groups. As discussed below, we cluster standard errors by study to account for dependence between these effect sizes.

Minority status was typically signaled via a foreign name and sometimes accent during phone inquiries; via name and other resume-related characteristics (e.g., foreign place of birth) in written applications; and via name, accent (if present), and physical appearance during the final interview stage. See Supplementary Material Appendix A for more details about the studies in our sample and how race is signaled at each stage.

Outcomes: The Discrimination Ratio and Difference from Callback to Job Offer

The basic measures of racial discrimination we compute for each study are discrimination ratios. This is the ratio of the percentage of callbacks or job offers received by white native-born applicants to the percentage of callbacks or job offers received by equally qualified applicants from an ethnic or racial minority group. Ratios above 1.0 indicate that native-born majority applicants received more positive responses than their comparable minority counterparts, with the amount above 1.0 multiplied by 100 indicating the relative size of this advantage.

Formally, let c_w be the number of callbacks received by native whites, and c_m be the number of callbacks received by the target minority groups (e.g., African Americans), and n_w be the number of applications submitted by white applicants, and n_m be the number of applications submitted by minority group members. The discrimination ratio for callbacks (y^c) is $\frac{c_w}{n_w} \div \frac{c_m}{n_m}$ or $\frac{c_w}{n_w} \times \frac{n_m}{c_m}$.[11]

We also create a similar discrimination ratio for job offers (y^j), where j is the number of job offers received by whites and minorities: $y^j = \frac{j_w}{n_w} \div \frac{j_m}{n_m}$ or $\frac{j_w}{n_w} \times \frac{n_m}{j_m}$. Because in-person audit studies match groups on their nonracial characteristics either through the assignment of characteristics or through random assignment, no further within-study controls are required for valid estimates of discrimination.

To assess the difference in discrimination between callback and job offer, we calculate the difference between each study's (logged) job offer and callback

Table 1. Job Offer Outcome Studies

Study ID	Country	Minority	Discrimination ratio, callback	Discrimination ratio, job offer conditional on callback	Discrimination ratio, job offer, unconditional	Log (job offer DR) – log (callback DR)	Sig.	Sig. (adj)	Initial applications/ inquiries
Allasino2004	Italy	Moroccan	1.962	1.292	2.535	0.256			1266
Arrijn1998	Belgium	Moroccan	1.871	1.279	2.394	0.246	*		2778
Attström2007	Sweden	Middle Eastern	2.046	0.923	1.888	−0.080			2646
Bendick1994	USA	African American	1.220	3.980	4.857	1.381	***	**	298
Bendick2010	USA	Tester Of Color	1.346	1.634	2.200	0.491			204
Bovenkerk1995	Netherlands	Moroccan	1.704	16.000	27.261	2.773	**	*	554
Cediey2008	France	North African	1.670	2.760	4.609	1.015	***	***	1392
Cediey2008	France	Sub-Saharan African	1.870	4.100	7.669	1.411	***	***	774
Cross1990	USA	Hispanic	1.331	1.140	1.518	0.131			720
Hjarno2008	Denmark	Pakistani	1.624	1.667	2.707	0.511			404
Hjarno2008	Denmark	Turkish	1.554	2.750	4.273	1.012	***	**	482
James1992	USA	Black	1.031	1.103	1.138	0.098			290
James1992	USA	Hispanic	1.071	0.597	0.640	−0.515			280
Prada1996	Spain	Moroccan	1.689	1.917	3.237	0.651	***	*	1104
Turner1991	USA	Black	1.144	1.306	1.494	0.267	**	*	952

Note: Twelve studies, fifteen effects (discrimination estimates against a distinct target group).
Sig. column gives significance of testing the null hypothesis that log(job offer DR) – log(callback DR) = 0.
Sig. (adj) gives significance tests that are Bonferroni-adjusted for fifteen comparisons.
N's for job offer conditional on callback column vary and are smaller than the number of initial applications.
*$p < .05$; **$p < .01$; ***$p < .001$, two-tailed tests.

discrimination ratios. If y^c_{im} is the callback discrimination ratio for minority group m in the ith study, and y^j_{im} is the job offer discrimination ratio for minority group m in the ith study, then the gap in discrimination between job offer and callback for minority group m in the ith study is $g_{im} = \ln(y^j_{im}) - \ln(y^c_{im})$. The ratios are logged to reduce the asymmetry for analysis purposes. Values greater than 0 indicate more discrimination overall in job offers received, and values less than 0 indicate more racial discrimination in callbacks received.

A statistic that is closely related to the difference between callback and job offer discrimination ratios is the job offer discrimination ratio conditional on callback($y^{j|c}_{im}$). One way to calculate this from the data is to calculate the job offer discrimination ratio only using data from auditors who received a callback: $y^{j|c}_{im} = \frac{j_w}{c_w} \div \frac{j_m}{c_m}$. Applying some algebra to this expression shows this is equivalent to the ratio of job offer to callback discrimination ratios, $y^{j|c}_{im} = y^j_{im} \div y^c_{im}$. Taking logs shows that the job offer discrimination ratio conditional on callback is equal to the difference in the logged discrimination ratios of job offer and callback (discussed in the previous paragraph): $\ln(y^{j|c}_{im}) = \ln(y^j_{im}) - \ln(y^c_{im}) = g_{im}$. The logged difference measure is more suitable for statistical analysis than the unlogged discrimination ratio conditional on callback because of its more symmetric distribution; however, the unlogged discrimination ratio conditional on callback is more interpretable; ultimately, these measures contain the same information presented in different ways.

Adjusting Discrimination Ratios in Studies with Conditional Following

A complication in our data is that some of the studies in our analysis use a conditional following rule for applicant pairs. In these studies, applicants only continue to the next stage if *both* members of the pair were successful during the prior stage: an applicant who received an invitation to interview, but whose audit partner did not, does not go to the interview. In these studies unconditional success rates (e.g., the unconditional rate of receiving job offers from first application) are not reported since not all auditors attempted to reach the later stages. Instead, the studies following this conditional rule only report each group's conditional success rates: success rates for each stage conditional on *joint* success of paired applicants at the prior stage.

We estimate the unconditional success rates by group at the callback and job offer stages by multiplying the unconditional probability of entering the stage by the conditional probability of success given entry to the stage. For example, we estimate the unconditional job offer rate for whites by multiplying the proportion of white applicants who receive a callback by the proportion of white applicants who receive a job offer conditional on receiving a callback.[12] We employ a similar method to estimate the unconditional job offer rates for minority applicants. This procedure requires the assumption that conditional group success rates in which both pair members receive a callback are the same as conditional group success rates in which only one pair member received a callback.[13]

Several of the studies using a conditional following rule also use a three-stage procedure. The stages are preliminary inquiry, in which auditors inquire by phone whether the advertised job is still available before applying; application conditional on successful pair preliminary inquiries; and attending the interview conditional on both members of a pair receiving callbacks.

In studies with conditional following, we estimate unconditional rates at each stage after the first stage from the success rate at the previous stage (actual or estimated) multiplied by the conditional success rate at the next stage. Further details of this procedure are given in Supplementary Material Appendix E. Supplementary Material Appendix F shows the conditional discrimination ratios at each stage in studies that use conditional following rules.

Variance Estimation

We calculate the variance of the discrimination ratio from counts of outcomes reported in each study using standard formulas for the variability of a ratio due to sampling error, accounting for audit pairs in the design when possible. For studies that are unpaired or do not report paired outcomes, the variance of the logged discrimination ratio for the mth minority group in the ith study for callbacks (y^c_{im}) is estimated by:

$$\sigma^2_{im} = Var\left(\ln\left(y^c_{im}\right)\right) = \frac{1}{c^w_{im}} - \frac{1}{n^w_{im}} + \frac{1}{c^m_{im}} - \frac{1}{n^m_{im}}$$

This is from Borenstein, Hedges, Higgins, and Rothstein (2009, formula 5.3). For studies that report paired results—with one minority and one white applicant applying for the same job—we use an alternative formula to account for the pairing (Zhou 2007, p. 27). Let p^a be the number of pairs in which both majority and minority testers receive a callback, p^b be the number of pairs in which the majority tester received a callback but not the minority, p^c be the number of pairs in which the minority tester received a callback but not the majority. The variance of the logged odds ratio for the mth minority group in the ith study with paired data is:

$$\sigma^2_{im} = Var\left(\ln\left(y^c_{im}\right)\right) = \frac{p^b_{im} + p^c_{im}}{\left(p^a_{im} + p^b_{im}\right)\left(p^a_{im} + p^c_{im}\right)}$$

We use the same formulas but substitute job offers for callbacks for the job offer outcome.

Our analysis focuses on the difference between the job offer discrimination ratio and the callback discrimination ratio. We calculate the variance of the difference in the logged callback and job offer discrimination ratios using estimates of the variability of the logged callback discrimination ratio and the logged job offer discrimination ratio (formulas above) and the estimated covariance of these statistics. We derived a formula for the covariance relying

on the fact that whenever a callback is not received (callback = 0) then by definition the job offer is not received (job offer = 0), creating positive covariance between job offer and callback outcomes (and therefore the job offer and callback discrimination ratios). This formula and other details are presented in Supplementary Material Appendix D. We calculate significance tests for the difference between the unconditional log job offer discrimination ratio and the log callback discrimination ratio for each study as a standard one-sample *t*-test of the null that the difference is zero, dividing the difference by its standard error to get the t-statistic. The standard error of the difference is also used in meta-analysis to estimate the average difference across studies.

Basic Data

Discrimination ratios for the twelve studies that make up our core sample, by target group, are shown in Table 1. As we can see, for callbacks the discrimination ratios ranged from 1.031 to 2.046, indicating that native whites receive 3.1% to 104.6% more callbacks than applicants from the minority group. For job offers, the discrimination ratios across studies range from 0.64 to 27.26, indicating that native whites receive 36% fewer job offers to twenty-seven times as many job offers as applicants from the minority group.[14] In only one case does the minority group receive more job offers than the majority: Latinos in the James and DelCastillo (1992) study.

Table 1 also shows the discrimination ratios for job offer conditional on callback ($y^{j|c}$). This is the discrimination ratio in job offers among respondents who successfully advanced to the callback stage. This conditional measure represents the additional discrimination that minority applicants face as they go from interview to job offer.[15]

By comparing the callback and job offer outcomes among applicants in the same study, study-level variables that have similar influences across stages are held constant and thus controlled. This includes many variables that could influence the outcome at the callback and job offer stages, like the qualifications given to applicants, the types of jobs applied for, and the broader social and national context. This method is equivalent to a model in which the outcomes for callback and job offer are separate effects but there is a fixed effect for study. Creating within-study ratios or difference measures is the typical way to do this in the meta-analysis literature.

Table 1 shows the difference in the logged job offer discrimination ratio and logged callback discrimination ratio and presents the test of significance for each minority group and study. For thirteen of the fifteen estimates, the job offer discrimination ratio is higher than the callback discrimination ratio. Eight of these fifteen differences are statistically significant at $p < .05$. Because we are looking across multiple tests, we also present *p*-values for the Bonferroni-adjusted tests (with fifteen tests) in the column next to the original *p*-values to guard against the possibility that some individual studies are only significant due to chance. Seven of the differences are still statistically significant after

Bonferroni adjustment. Overall, the results in Table 1 support the conclusion of more discrimination in job offers than in callbacks.[16]

Meta-Analysis Statistical Model

Although the statistics in Table 1 indicate higher discrimination in job offers than in callbacks for many studies, they do not provide an estimate of the overall magnitude of this difference. To provide an overall best estimate, we use techniques from the meta-analysis literature to combine data to create a point estimate of the magnitude of the difference across studies.

Specifically, we use a random-effects meta-analysis model to analyze the differences between job offer and callback outcomes. Random effects incorporate a variance component capturing unexplained variation in outcomes across studies (Raudenbush 2009). Random effects are recommended whenever there is reason to believe that the effect in question varies as a result of study-level variables. This is the case in our analysis, as we expect that the level of racial discrimination may depend on the year of the study or the situation the study considers (e.g., the country), the methodology of the study, and so on. The random effect increases the standard errors of estimates to account for random variation in outcomes due to study-level factors.

More formally, random-effects meta-analysis allows the true gap between callback and job offer discrimination ratios to be estimated on average across studies by assuming that study gaps have a normal distribution around the population mean gap between callbacks and job offers, θ. If g_{im} is the gap in the logged discrimination ratio between callback and job offer for the mth minority group in the ith study, then the meta-analysis model is:

$$g_{im} = \theta + u_i + e_{im}, \text{ where } u_i \sim N(0, \tau^2) \text{ and } e_{im} \sim N(0, \sigma_{im}^2).$$

There are no predictor variables in this model, just an average effect and a random effect for study-level variation. Here τ^2 is the between-study variance, estimated as part of the meta-analysis model, while σ_{im}^2 is the variance of the logged discrimination ratio of the mth minority group in the ith study, estimated from study outcome counts as described above. The between-study variance is estimated from the residual variation in study outcomes not accounted for by random sampling variation; estimation is by restricted maximum likelihood (see Raudenbush 2009). In computing the average effect, each study estimate is assigned a weight proportional to its inverse variance, $1/(\tau^2 + \sigma_{im}^2)$. We also perform meta-analyses where the outcome is the logged callback or job offer discrimination ratio, $\ln(y^c)$ or $\ln(y^j)$, respectively.

In one analysis, we add controls for some study characteristics to predict the difference between job offer and callback discrimination ratios. We do this with meta-regression (Raudenbush 2009; Borenstein et al. 2009). Meta-regression allows us to model the difference in the logged callback and job offer discrimination ratios as a function of a vector of k characteristics of the studies

and effects, x, plus residual study-level heterogeneity (between-study variance not explained by the covariates, the "random effect" in the meta-regression).

The model assumes the study-level heterogeneity follows a normal distribution around the linear predictor:

$$g_{im} = x_{im}\beta + u_i + e_{im}, \text{where } u_i \sim N(0, \tau^2) \text{ and } e_{im} \sim N(0, \sigma^2_{im})$$

where β is a k × 1 vector of coefficients (including a constant), and x_{im} is a 1 × k vector of covariate values for minority group m in study i (including a 1 for a constant).

Small Sample Adjustments and Accounting for Dependence of Discrimination Estimates

Our basic analysis is based on twelve studies. Sample sizes in this range are common in meta-analysis (e.g., Brockwell and Gordon 2001). We employ small-sample corrections with clustered standard errors—see Tipton (2015)—to help account for the effects of small sample sizes on inferential statistics.

Some studies estimate discrimination against more than one target group, for instance, blacks and Latinos (see Table 1 for a list of studies and groups). This is why we have fifteen discrimination estimates ("effects") based on the twelve studies. The use of multiple groups from the same study, based on many similar procedures and sometimes a common majority control group, creates dependence among the estimates that must be adjusted for when calculating inferential statistics. To do this, we cluster standard errors at the study level, allowing for dependence of effects in the same study. Following procedures suggested by Tipton (2015), we estimate results using "correlated" weights with an assumed correlation of 0.8. Estimation is done with the "robumeta" command in Stata and R (Fisher and Tipton 2015).[17]

Publication Bias

A potential problem in meta-analysis (and other methods of reviewing literature) is publication bias, or the concern that studies with null effects might be less likely to be published, resulting in an upward bias of effect estimates. In meta-analysis, a series of tests that are potentially diagnostic of publication bias are based on checking for a correlation between study sample size and effect size (Sutton 2009).

We examined the results of two standard tests of publication bias for the job offer outcomes, the Egger test and the trim-and-fill procedure (see Sutton 2009). The Egger test gave a borderline p-value of 0.075; the trim-and-fill tests suggest no points need to be imputed, suggesting no publication bias. If we drop the outlier of the Bovenkerk, Gras and Ramsoedh's (1995) study (see endnote 19) then the Egger test p-value is 0.550; the trim-and-fill test still suggests no observations should be imputed. We conclude from this there is no significant evidence of publication bias.

Figure 2. Discrimination ratios by stage.

Note: Dots indicate point estimates of mean discrimination ratios from meta-analysis shown in Table 2 Panel A. Lines indicate 95% confidence intervals. Dotted red line is discrimination ratio of 1.0 (no discrimination).

A lack of publication bias is not too surprising in the case of audit studies that go to the job offer outcome, because the studies are sufficiently difficult to conduct that authors are likely to produce a publication or public report regardless of whether or not they found evidence of discrimination. Most of the studies are funded or sponsored by large organizations (e.g., the International Labor Organization) that require reports to be produced regardless of outcome. Moreover, given the existence of many field experiments that find evidence of discrimination in hiring (Neumark 2018; Quillian et al. 2019), null findings may be viewed as novel and therefore publishable.

Results

We begin with a basic meta-analysis of the level of discrimination at different stages. Results of the meta-analysis for each stage are shown in Figure 2 and Table 2 Panels A and B. For our sample of twelve studies, the results indicate that majority applicants receive 53% more callbacks than equally qualified minority applicants on average (discrimination ratio of 1.534; 95% confidence interval of 1.33–1.78).

What happens after the callback? The discrimination ratio for job offers conditional on receiving a callback (i.e., only for applicants who made it to the interview stage) is 1.534[18]; this indicates that even when both candidates receive an interview, majority applicants still receive about 50% more job offers than comparable minority applicants. Looking at the overall level of

Table 2. Meta-Analysis of Callback and Job Offer Outcomes

A. Stage of Hiring ($N = 12$ studies, 15 effects)	Mean discrimination ratio	Sig.	95% CI		Tau-squared
Callback	1.534	***	1.325	1.775	0.0451
Job offer conditional on callback	1.534	*	1.126	2.088	0.1033
Job offer unconditional	2.450	***	1.682	3.568	0.1819
B. Differences between stages (N = 12 studies, 15 effects)	**Mean difference in log discrimination ratios**				
Callback and job offer conditional on callback	0.016		−0.331	0.362	0.1466
Callback and job offer, unconditional	0.428	*	0.119	0.736	0.1033
C. Comparison to field experiments with callback outcome only (75 studies, 111 effects)	**Mean discrimination ratio or difference in log ratios**				
Callback, field experiments with callback outcome only	1.537	***	1.421	1.662	0.0775
Log difference callback outcome (75 studies) and job offer unconditional (12 studies)	0.423	*	0.066	0.780	0.0816

*$p < .05$; **$p < .01$; ***$p < .001$. Two-tailed tests.

Note: Random-effects meta-analysis with clustered standard errors at the study level, using "correlated" cluster weights (Tipton 2015) with assumed rho = .8. Mean discrimination ratios in panels A and C are based on the anti-log of the mean logged discrimination ratio. Tau-squared is the estimated between-study variance in log discrimination ratios. CI, confidence interval.

discrimination in job offers, majority applicants receive about 145% more job offers than comparable minority applicants (discrimination ratio of 2.450, 95% confidence interval of 1.68–3.57).[19] The difference between the callback discrimination ratio and the unconditional (or overall) job offer discrimination ratio is statistically significant at $p < 0.05$ (shown in Panel B of Table 2). These results indicate that there is a considerable degree of additional discrimination against racial minorities as they move from callback to job offer. The point estimates suggest that minority candidates experience on average more than twice as much discrimination overall in the job offer outcome as in the callback outcome.

Figure 3 shows a forest plot of the difference in the logged discrimination ratios for the callback and job offer outcomes for each of the fifteen effect sizes (based on the twelve studies) in our analysis. For thirteen of the fifteen effect sizes, the point estimate is greater than zero, which indicates a greater discrimination ratio for job offers than for callbacks; the two effect sizes with less discrimination in job offers than in callbacks are not statistically significant differences (at $p < 0.05$). There is thus a fairly consistent pattern across studies of greater discrimination in job offers than in callbacks.

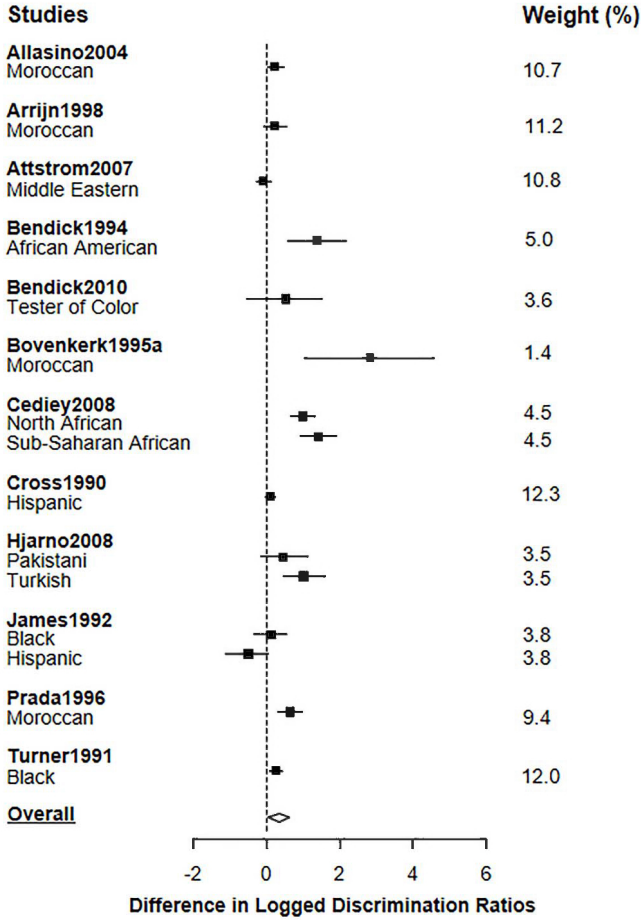
How generalizable are our findings based on twelve studies that go to the job offer stage to the broader population of field experiments of racial discrimination in hiring? To address this question, we compare the average callback discrimination ratio of the twelve studies in our sample that went to the job offer with the average callback discrimination ratio for all field experimental studies located in our meta-analysis search (see Supplementary Material Appendix B) that were conducted in the same countries as the job offer outcome studies.

Results are shown in Table 2 Panel C. The average discrimination ratio in the callback outcome studies ($n = 75$ studies) is 1.537. The average callback discrimination ratio in the job offer studies ($n = 12$ studies) is 1.534 (panel A). This is not a statistically significant difference, providing some evidence that it is reasonable to draw conclusions about field experiments of hiring in general from the job offer studies.

Next, we use meta-regressions to address two related questions. First, does the level of discrimination at the callback stage predict the level of discrimination in job offers? Most studies use callbacks to draw conclusions about racial discrimination in hiring. We would expect some association since failure to receive a callback generally means that there is no chance of receiving a job offer. Figure 4 graphs the line of best fit in predicting the logged unconditional (or overall) job offer discrimination ratio as a function of the logged callback discrimination ratio. The Pearson correlation of the two stages (scatterplot in Figure 4) is 0.55. In summary, studies that find high callback discrimination also tend to find high job offer discrimination.

Second, does the level of discrimination at the callback stage predict the magnitude of the additional discrimination that occurs at the job offer stage? The line of best fit in predicting conditional job offer discrimination from callback discrimination is shown in Figure 5. The line is fairly flat. The Pearson correlation of the two stages (scatterplot in Figure 5) is 0.21. The level of

Figure 3. Forest plot of difference in discrimination ratios job minus callback.

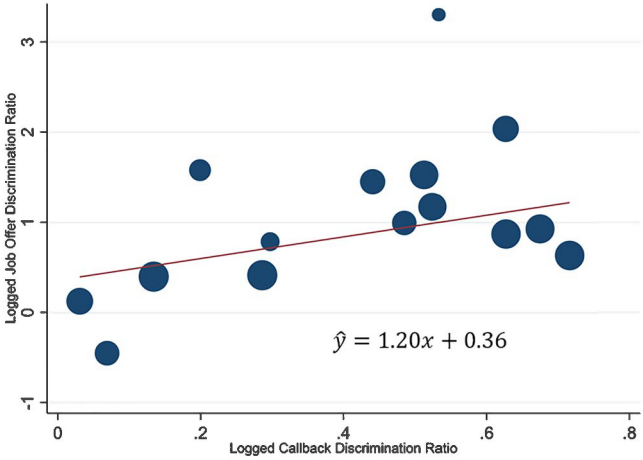


Note: Squares are point estimates and lines are 95% confidence intervals. Diamond in overall column is 95% confidence interval for the overall effects. Weight indicates importance in determining overall effect.

discrimination at the callback stage is only weakly correlated with the *additional* discrimination that occurs after the callback (i.e., conditional on having received an invitation to interview). That is, there is a considerable degree of heterogeneity with respect to how racial prejudices against minority candidates unfold during the hiring process.

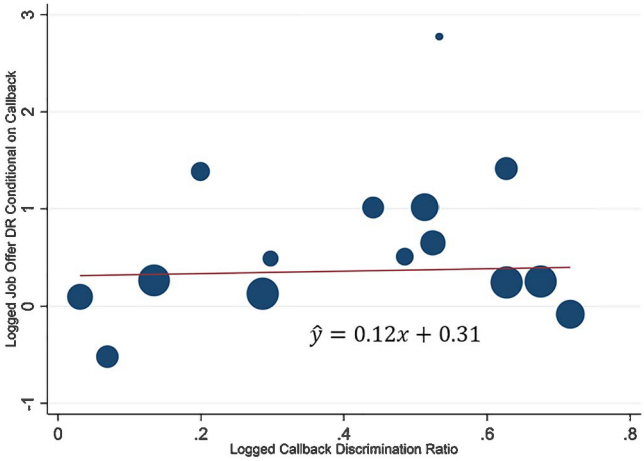
Finally, we examine whether study-level factors are associated with the magnitude of the difference in discrimination between the callback and job offer stages. We do this by estimating a meta-regression in which the difference in the (logged) discrimination ratios between callback and job offer is the outcome as a function of national, group, and time covariates. Because there are only twelve

Figure 4. Meta-regression of unconditional job offer discrimination ratio on callback discrimination ratio.



Note: Line from meta-regression. Marker sizes are proportional to random effect weights.

Figure 5. Meta-regression of conditional job offer discrimination ratio on callback discrimination ratio.



Note: Line from meta-regression. Marker sizes are proportional to random effect weight.

studies, we can only include a few predictors, and statistical power is low. Only large effects will be statistically significant.

Table 3 shows models that include a dummy variable for the United States vs. Europe, dummy variables for black/African and Middle-Eastern/North African target groups, and a control for year of the study. Either alone or with other controls, none of the coefficients of these predictors are significant at conventional levels. The only regressor that is close to significant ($p < 0.20$) both alone

Table 3. Meta-Regression of Difference in Job Offer and Callback Discrimination Ratios

Outcome: Difference of log job offer and log callback discrimination ratios				
	(1)	(2)	(3)	(4)
Country = United States (vs. Europe)	−0.23			−0.70
	(0.28)			(0.42)
Minority Group Black/African (1 = yes)		0.56		0.54
		(0.35)		(0.34)
Minority Group Middle-Eastern/North African (1 = yes)		0.33		−0.23
		(0.24)		(0.35)
Year of fieldwork of study (four digit year)			0.01	0.00
			(0.03)	(0.03)
Constant	0.48	0.15	−17.85	7.72
	(0.19)	(0.16)	(51.27)	(58.39)
Tau-squared	0.1213	0.1799	0.1237	0.1678
N studies/N effects	12/15	12/15	12/15	12/15

Note: Random-effects meta-regression with clustered standard errors at the study level, using "correlated" cluster weights (Tipton 2015) with assumed rho = .8. Standard errors in parentheses.

and with controls is black/African. We thus have some evidence there might be a greater disparity in discrimination between callbacks and job offers for black/African applicants.

Do In-Person Audits Overestimate Discrimination?

Finally, we consider a potential methodological problem with in-person testing raised in the literature. In a critique of in-person audit studies, Heckman and Siegelman (1993) argued that systematic bias may tend to enter in-person audit studies because in most studies the auditors know the purpose of the experiment is to detect discrimination.[20] They suggest that auditors may (consciously or not) tend to act in accord with this purpose, resulting in auditors being more likely to find discrimination. For example, minority auditors may perform poorly in the interview, reflecting a desire to "find" discrimination, resulting in fewer job offers for minority candidates and an overestimate of actual discrimination.

To the best of our knowledge, no evidence supports Heckman and Siegelman's argument. Nevertheless, it remains an important possibility, and this possibility is one factor that has motivated the use of resume audits (Cherry and Bendick 2018).

One previous study provides evidence against the idea that experimenter effects in in-person studies contribute to racial disparities in job outcomes: Pager (2007) conducted an in-person audit for jobs with a callback outcome. As part of her procedure, applicants recorded whether or not they had significant contact with the employer or person in charge of hiring (applicants were instructed to request the person in charge of hiring when they arrived). In many cases, applicants had little contact with the employer or hiring agent because that person was not there or too busy, and merely submitted an application. In other cases, applicants had significant contact with the hiring agent, either in the form of an on-the-spot interview or a conversation about the job. Pager found that for both white and black auditors, having face-to-face contact with the employer significantly increased the likelihood of receiving a callback, with a larger boost for black auditors (Pager 2007, chapter 4 appendix A, chapter 6).[21] The fact that face-to-face contact increased the likelihood of a positive response and decreased the racial gap provides some evidence against the problem of experimenter effects. However, it would be desirable if future research using in-person audits could be conducted double-blind to address this complaint directly.

Finally, we note that features of field experimental study design may result in understating the true levels of discrimination faced by minorities in face-to-face interactions. Investigators train auditors and match them in pairs to provide similar self-presentation styles, generally following the cultural patterns and norms of the majority (white) group. For instance, auditors lack strong accents and usually dress in standard business attire. Because of this, in face-to-face field experiments the ethnic markers of minority candidates tend to be muted. However, many minority applicants in the real labor market who are otherwise qualified do have culturally distinct norms and styles of dress. Field experiments may thus understate discrimination because they tend to suppress the display of ethnic cultural characteristics that may operate as the basis of discrimination.

Discussion

Our meta-analysis of studies that go to the job offer indicates that racial discrimination in hiring is substantially more severe than an analysis of solely callback outcomes would suggest. Majority applicants in our sample receive 53% more callbacks than comparable minority applicants, but they receive 145% more job offers than comparable minority applicants. The higher level of discrimination in job offers is a fairly consistent result: in thirteen of our fifteen estimates of discrimination against minority groups there is more discrimination in job offers than in callbacks. The job offer outcome represents the accumulation of discrimination from application to callback and from interview to job offer, and there is substantial, additional discrimination at the second stage.

To get a sense of the impact of this finding on the apparent level of discrimination in labor markets, consider a recent meta-analysis of callback studies by Quillian et al. (2017). The authors found that on average, white applicants in the United States received 36% more callbacks than comparable black applicants,

and 24% more callbacks than comparable Latino applicants. While this racial discrimination in callbacks seems like a serious problem, it appears to be much more serious if white applicants are actually receiving 72% more job offers than equally qualified black applicants, and 48% more job offers than equally qualified Latino applicants—statistics somewhat below those suggested by our point estimates.[22]

What do these results indicate about the widespread use of callbacks in studies as a proxy outcome for hiring discrimination? Callbacks tend to understate the total level of discrimination in hiring, but callbacks are a reasonable proxy in a *relative* sense: that is, studies finding high levels of discrimination in callbacks also generally find high levels of discrimination in job offers. This is because close to half of the total discrimination in hiring occurs from initial application to callback, and because the additional discrimination from interview to job offer is weakly positively correlated with the level of discrimination in callbacks (and thus does not tend to offset earlier discrimination, as it would if later discrimination were negatively correlated with earlier discrimination).

What do these results imply about the appropriate outcome measure in future field experiments of hiring discrimination? Because job offer studies are much more difficult to conduct than callback studies and also involve much greater ethical concerns, we think it is impractical to entirely replace callback studies with job offer studies.

However, we think that it makes sense to continue some job offer outcome studies in cases where resources are available and to test assumptions about how callbacks and job offers correspond. For example, the most recent job offer study in the United States is by Bendick et al. (2010), which only sent testers to apply for positions as waiters and waitresses in New York City. The last U.S. study with a job offer outcome that sent testers to apply for a wider variety of positions was published during the early 1990s. How has the connection between callback discrimination and job offer discrimination changed (or not changed) over the past 20 years in the United States? We lack evidence to answer this question. In this context, we believe a new job offer study for the near future in the United States labor market could help clarify how discrimination post-callback has changed in the United States over the last two decades.

Our results also have implications regarding the nature of racial discrimination in hiring. Discrimination by employers does not appear to function in a categorical way, in which employers who know the race or ethnicity of an applicant pre-callback automatically rule out minority applicants in favor of equally qualified majority applicants. Instead, racial discrimination in hiring has a probabilistic character across stages of hiring, in which minority applicants are less likely to advance at each stage.

Our results also provide evidence against the view that most racial discrimination in hiring reflects statistical discrimination, at least on the characteristics employers can assess during an interview. If the main reason employers discriminate is statistical, employers should be less likely to rely on group stereotypes in drawing conclusions about applicants as their information about an individual applicant increases. In interviews, employers receive additional information

regarding (at least) the dress, appearance and demeanor of applicants. Yet we find as much discrimination or more from interview to job offer as from application to callback, suggesting that the extra interview information is not significantly counter-acting employer biases.

Our meta-analysis also has some limitations arising from the sample of studies on which it is based. First, the studies we use are from a variety of social contexts, including many countries and several different target groups. Unfortunately, we do not have enough studies to estimate with precision how social context affects differences in the levels of discrimination between callback and job offer—most of our results regarding such factors are inconclusive. Second, our confidence intervals are somewhat wide. This reflects the fact that there are only twelve recent field experimental studies of racial discrimination in hiring that meet our inclusion criteria and assess the job offer outcome. However, our key results are statistically significant at conventional levels even given this small sample. In addition, the average callback result for the twelve studies we analyzed look very similar to that of the seventy-five field experiments in these same countries that stopped at the callback, suggesting that the twelve studies that go to the job offer are not atypical of audit studies more generally. Third, while we show that there is substantial additional discrimination at the job offer stage, we cannot ascertain the specific reason for this discrimination from our results.

Our results show that minority applicants face substantial additional discrimination even after they make it past the initial resume selection process. Racial discrimination in the labor market is thus significantly more serious than is suggested by field experiments that stop with the callback. Although our results pertain only to racial and ethnic discrimination in hiring, gaps between callback and job offer outcomes may also manifest for other bases of discrimination, such as age, gender, religion or sexual orientation. For discrimination on other bases, the correspondence between callback and job offer discrimination has yet to be tested. In light of the high social and economic costs of discrimination in hiring, this is another potentially fruitful area for future research.

Supplementary Material

Supplementary material is available at *Social Forces* online, http://sf.oxfordjournals.org/.

Notes

1. Authors' count; see "Methods" and Supplementary Material Appendix B for details of how the search was conducted.
2. Author's count. Of the more than one hundred field experiments we have identified, only thirteen pursue audits to the job offer outcome.
3. See Supplementary Material Appendix A for a list of the methods used in the twelve in-person audit studies included in our sample to signal race or ethnicity. Only two estimates of discrimination (of the fifteen) did not use

name as a signal; both assess discrimination against African Americans based on in-person applications.

4. Not all face-to-face studies go to the job offer outcome; some use in-person applications but focus on receiving callbacks and do not have auditors return for interviews if invited back (e.g., Pager, Bonikowski and Western 2009).
5. Gaddis (2017a) finds that only 19% of African Americans in New York have racially distinctive names. For many African-American applicants, then, race may not play a role until the interview stage. However, racially distinctive names are more common among immigrant minorities. Also in many lower-pay occupations (e.g., food service and retail) in-person applications are still common.
6. Kang et al. (2016) find that some minority candidates attempt to "whiten" their resumes by reducing racial cues; in such cases, the candidates' race would not be apparent until the interview.
7. However, in-person audit studies generally try to match auditors on interactional style, so this is likely less important in creating discrimination in the audit context.
8. Another possibility is that employers discriminate less in the callback stage in a conscious effort to avoid being caught by government regulators conducting audits. Evidence supports this occurring in the housing sector (Ross and Turner 2005; Galster and Ross 2007). In the US employment sector this seems less likely because the EEOC is banned by administrative rules from using audits.
9. A distinct point discussed in the field experimental literature is whether it makes more sense to calculate measures of discrimination by including or excluding paired tests in which neither tester gets a callback (e.g., Neumark 2018, p. 825).
10. A bibliographic list of these studies can be found in Supplementary Material Appendix G.
11. Two other measures that could be used instead of the ratio of callback/job offer percentages are the odds ratio and the difference in proportions. We prefer the discrimination ratio for reasons discussed in Supplementary Material Appendix C. This appendix also discusses results using these alternative measures.
12. This is applying the definition of conditional probability: probability(B) = probability(B|A) × probability(A).
13. It might seem tempting to code successful applicants who do not go to the next stage (because their partner was not successful) as not getting a job offer. However, this would underestimate success rates because some testers who do not continue would receive job offers if they continued the application process. Moreover, underestimates will tend to be larger for the majority group because the majority group more often receives positive callbacks.
14. The Bovenkerk1995 study that finds a discrimination ratio of 27.26 in job offers is an outlier. In this study none of the minority auditors received job offers; we assigned ½ job offer to compute the discrimination ratio. Dropping this study has no effect on our results, see footnote 19.

15. Conditioning on callback in calculating job offers conditions on a post-treatment variable, which can create bias in analyses of experiments (Coppock 2019; Montgomery et al. 2018). If the goal of our analysis is to estimate discrimination in job offers, it would be a serious problem to only use the post-callback estimate of discrimination (since it conditions out discrimination at the callback stage). However, since our goal is to contrast callback and job offer outcomes, job offers conditional on callbacks provide a clear way to understand the additional discrimination that occurs post-callback.
16. Looking at the results jointly, the chance of eight or more tests being significant out of fifteen if there are no significant differences is less than .0001. This probability can be calculated using the binomial formula (see Ross et al. 2008, section 4.1).
17. Using "hierarchical" weights (Tipton 2015) also produced similar results.
18. It is a coincidence, not an error, that the job offer discrimination ratio conditional on callback is the same as the callback discrimination ratio rounded to three digits.
19. If we drop the Bovenkerk, Gras and Ramsoedh 1995 study, which has a very high job offer discrimination ratio (27.26), this decreases the meta-analysis estimated average job offer discrimination ratio from 2.450 to 2.320, leaving our conclusions unchanged.
20. National Research Council (2004) discusses Heckman and Siegelman's other critiques. Neumark (2012) has also argued that audits may reflect employer perceptions of different variability in job-relevant characteristics across groups. However, this would be a type of statistical discrimination, and for this reason, still illegal in that it involves judging individuals based on their group membership.
21. Without criminal backgrounds.
22. We also find that the difference in discrimination between the callback and job offer could be somewhat larger when the target group is black; see the discussion of Table 3. If this is the case, then relative to black or African Americans, the advantage in job offers enjoyed by whites could be larger than 72%.

About the Authors

Lincoln Quillian is Professor of Sociology and Fellow of the Institute for Policy Research at Northwestern University. His past work includes studies of neighborhood poverty concentration, internal migration, racial residential segregation and racial attitudes. His recent research analyzes results from audit and correspondence studies to better understand racial and ethnic discrimination in hiring around the world.

John J. Lee is a PhD candidate in sociology at Northwestern University. His research interests include social psychology, inequality, race/ethnicity and public opinion.

Mariana Oliver is a joint JD/PhD student in sociology at Northwestern University, and a Law and Science Fellow at Northwestern Pritzker School of Law. This article is part of her broader interest in how organizational design and policies matter for various forms of inequality. Her dissertation examines the use of surveillance tools and technology in police departments and its implications for privacy rights.

References

Agerström, Jens, Fredrik Björklund, Rickard Carlsson, and Dan-Olof Rooth. 2012. "Warm and Competent Hassan = Cold and Incompetent Eric: A Harsh Equation of Real-Life Hiring Discrimination." *Basic and Applied Social Psychology* 34(4):359–66.

Altonji, Joseph G., and Charles R. Pierret. 2001. "Employer Learning and Statistical Discrimination." *The Quarterly Journal of Economics* 116(1):313–350.

Arrow, Kenneth 1973. "The Theory of Discrimination." In *Discrimination in Labor Markets*, edited by Ashenfelter, Orley C., Rees, Albert, pp. 3–33. Princeton: Princeton University Press.

Attström, Karin 2007. "Discrimination against Native Swedes of Immigrant Origin in Access to Employment." In *86E. International Migration Papers*. Geneva, Switzerland: International Labour Organization.

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 Gaddis, S. Michael, pp. 63–77. Methodos. Berlin, Germany: Springer.

Bendick, Marc, Rekha Eanni Rodriguez, and Sarumathi Jayaraman. 2010. "Employment Discrimination in Upscale Restaurants: Evidence from Matched Pair Testing." *The Social Science Journal* 47(4):802–818.

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.

Bobo, Lawrence, Camille Charles, Maria Krysan and Alicia Simmons 2012. "The Real Record on Racial Attitudes." In *Social Trends in American Life*, edited by Marsden, Peter V., pp. 38–83. Princeton, NJ: Princeton University Press.

Borenstein, Michael, Larry V. Hedges, Julian P.T. Higgins and Hannah R. Rothstein 2009. *Introduction to Meta-Analysis*. West Sussex, UK: John Wiley & Sons.

Bovenkerk, F., M.J.I. Gras and D. Ramsoedh 1995. "Discrimination against Migrant Workers and Ethnic Minorities in Access to Employment in the Netherlands." In *International Migration Papers*. Geneva, Switzerland: International Labour Organization.

Brockwell, Sarah E., and Ian R. Gordon. 2001. "A Comparison of Statistical Methods for Meta-Analysis." *Statistics in Medicine* 20(6):825–840.

Bursell, Moa. 2014. "The Multiple Burdens of Foreign-Named Men—Evidence from a Field Experiment on Gendered Ethnic Hiring Discrimination in Sweden." *European Sociological Review* 30(3):399–409.

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.

Cediey, E. and F. Foroni 2008. "Discrimination in Access to Employment on Grounds of Foreign Origin in France." In *85E. International Migration Papers*. Geneva, Switzerland: International Labour Organization.

Cherry, Frances and Marc Bendick 2018. "Making It Count: Discrimination Auditing and the Activist Scholar Tradition." In *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*, edited by Gaddis, S. Michael. Methodos Series. Cham, Switzerland: Springer International Publishing AG.

Coppock, Alexander. 2019. "Avoiding Post-Treatment Bias in Audit Experiments." *Journal of Experimental Political Science* 6(1):1–4.

Crabtree, Charles and Volha Chykina 2018. "Last Name Selection in Audit Studies." *Sociological Science* 5:21–8.

Fisher, Zachary, and Elizabeth Tipton. 2015. *Robumeta: An R Package for Robust Variance Estimation in Meta-Analysis*. https://arxiv.org/abs/1503.02220.

Gaddis, Michael. 2017a. "How Black Are Lakisha and Jamal? Racial Perceptions from Names Used in Correspondence Audit Studies." *Sociological Science* 4(19):469–89.

Gaddis, Michael 2017b. "Racial/Ethnic Perceptions from Hispanic Names: Selecting Names to Test for Discrimination." *Socius: Sociological Research for a Dynamic World* 3:1–11.

Gaddis, Michael. 2015. "Discrimination in the Credential Society: An Audit Study of Race and College Selectivity in the Labor Market." *Social Forces* 93(4):1451–1479.

Gaddis, Michael (ed.) 2018. *Audit Studies: Behind the Scenes with Theory, Method, and Nuance. Vol. 14. Methodos Series*. Cham, Switzerland: Springer International Publishing AG.

Galster, George and Stephen L. Ross 2007. "Fair Housing Enforcement and Changes in Discrimination between 1989 and 2000." In *Fragile Rights within Cities: Government, Housing, and Fairness*, edited by Goering, John. Lanham, MD: Rowman & Littlefield.

Gerber, Alan S. and Donald P. Green 2012. *Field Experiments: Design, Analysis, and Interpretation*, 1st edn. New York: W. W. Norton.

Greenwald, Anthony G., Debbie E. McGhee, and Jordan L. K. Schwartz. 1998. "Measuring Individual Differences in Implicit Cognition: The Implicit Association Test." *Journal of Personality and Social Psychology* 74(6):1464–80.

Guryan, Jonathan, and Kerwin Kofi Charles. 2013. "Taste-Based or Statistical Discrimination: The Economics of Discrimination Returns to its Roots." *The Economic Journal* 123(572): F417–32.

Heckman, James J. and Peter Siegelman 1993. "The Urban Institute Audit Studies: Their Methods and Finding." In *Clear and Convincing Evidence: Measurement of Discrimination in America*, edited by Fix, Michael, Struy, Raymond J., pp. 187–258. Lanham, MD: Urban Institute Press.

James, Franklin J. and Steven W. DelCastillo 1992. "Measuring Job Discrimination: Hopeful Evidence from Recent Audits." *Harvard Journal of African American Public Policy* 1:33–53.

Kang, Sonia K., Katherine A. DeCelles, András Tilcsik, and Sora Jun. 2016. "Whitened Résumés: Race and Self-Presentation in the Labor Market." *Administrative Science Quarterly* 61(3):469–502.

Kirschenman, Joleen and Kathryn M. Neckerman 1991. "We'd Love to Hire Them, but . . . ' : The Meaning of Race for Employers." In *The Urban Underclass*, edited by Jencks, Christopher, Peterson, Paul E., pp. 203–32. Washington, DC: Brookings.

Montgomery, Jacob M., Brendan Nyhan, and Michelle Torres. 2018. "How Conditioning on Posttreatment Variables Can Ruin your Experiment and What to Do about it." *American Journal of Political Science* 62(3):760–75.

National Research Council 2004. *Measuring Racial Discrimination*. Washington, D.C.: The National Academies Press.

Neumark, David. 2012. "Detecting Discrimination in Audit and Correspondence Studies." *Journal of Human Resources* 47(4):1128–57.

Neumark, David. 2018. "Experimental Research on Labor Market Discrimination." *Journal of Economic Literature* 56(3):799–866.

Pager, Devah, Bart Bonikowski, and Bruce Western. 2009. "Discrimination in a Low-Wage Labor Market: A Field Experiment." *American Sociological Review* 74(5):777–99.

Pager, Devah, and Lincoln Quillian. 2005. "Walking the Talk? What Employers Say Versus What they Do." *American Sociological Review* 70(3):355–80.

Pager, Devah 2007. *Marked: Race, Crime, and Finding Work in an Era of Mass Incarceration*. Chicago: University of Chicago Press.

Pager, Devah and Hana Shepherd 2008. "The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets." *Annual Review of Sociology* 34:181–209. doi: 10.1146/annurev.soc.33.040406.131740.

Pascoe, Elizabeth A., and Laura Smart Richman. 2009. "Perceived Discrimination and Health: A Meta-Analytic Review." *Psychological Bulletin* 135(4):531–54.

Quillian, Lincoln, Anthony Heath, Devah Pager, Arnfinn Midtbøen, Fenella Fleischman and Ole Hexel 2019. "Do some Countries Discriminate More than Others? Evidence from 97 Field Experiments of Racial Discrimination in Hiring." *Sociological Science* 6:467–96.

Quillian, Lincoln, and Devah Pager. 2010. "Estimating Risk: Stereotype Amplification and the Perceived Risk of Criminal Victimization." *Social Psychology Quarterly* 73(1):79–104.

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):10870–5.

Raudenbush, Stephen W. 2009. "Adaptive Centering with Random Effects: An Alternative to the Fixed Effects Model for Studying Time-Varying Treatments in School Settings." *Education Finance and Policy* 4(4):468–91.

Rivera, Lauren A. 2012. "Hiring as Cultural Matching: The Case of Elite Professional Service Firms." *American Sociological Review* 77(6):999–1022. https://doi.org/10.1177/0003122412463213.

Rooth, Dan-Olof. 2010. "Automatic Associations and Discrimination in Hiring: Real World Evidence." *Labour Economics* 17(3):523–34. doi:10.1016/j.labeco.2009.04.005.

Ross, Stephen L., and Margery Austin Turner. 2005. "Housing Discrimination in Metropolitan America: Explaining Changes between 1989 and 2000." *Social Problems* 52(2):152–180.

Ross, Stephen L., Margery Austin Turner, Erin Godfrey, and Robin R. Smith. 2008. "Mortgage Lending in Chicago and Los Angeles: A Paired Testing Study of the Pre-Application Process." *Journal of Urban Economics* 63(3):902–19.

Sutton, Alex J. 2009. "Publication Bias." In *The Handbook of Research Synthesis and Meta-Analysis*, edited by Cooper, Harris, Hedges, Larry V., Valentine, Jeffrey C., 2nd edn. New York, NY: Russell Sage Foundation.

Tipton, Elizabeth. 2015. "Small Sample Adjustments for Robust Variance Estimation with Meta-Regression." *Psychological Methods* 20(3):375–93.

Weichselbaumer, Doris 2016. "Discrimination against Female Migrants Wearing Headscarves." IZA Discussion Paper No. 10217. http://ftp.iza.org/dp10217.pdf.

Williams, David, and Pamela B. Jackson. 2005. "Social Sources of Health Disparities". *Health Affairs* 24:325–34.

Zegers de Beijl, Roger. 2000. *Documenting Discrimination Against Migrant Workers in the Labor Market: A Comparative Study of Four European Countries*. Geneva, Switzerland: International Labour Organization.

Zhou, Guang Yong. 2007. "One Relative Risk Versus Two Odds Ratios: Implications for Meta-Analyses Involving Paired and Unpaired Binary Data." *Clinical Trials* 4:25–31.

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.