Northwestern | INSTITUTE FOR POLICY RESEARCH

Working Paper Series

WP-18-28

Meta-Analysis of Field Experiments Shows Significantly More Racial Discrimination in Job Offers than in Callbacks

Lincoln Quillian
Professor of Sociology and IPR Fellow
Northwestern University

John Lee
Graduate Student in Sociology
Northwestern University

Mariana Oliver
Graduate Student in Sociology
Northwestern University

Version: November 6, 2018

DRAFT
Please do not quote or distribute without permission.

ABSTRACT

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 racial discrimination in the hiring process based on an assessment of differences in callback rates, when the ultimate subject of interest is discrimination in job offers? To address this question, the researchers perform a statistical meta-analysis of all available field experimental studies of racial discrimination in hiring that go to the job offer outcome. Their sample includes 12 studies encompassing more than 8,300 job applications. They find significant additional discrimination in hiring after the callback: Majority applicants in our sample receive 52% more callbacks than comparable minority applicants, but majority applicants receive 128% more job offers than comparable minority applicants. The additional discrimination from interview to job offer is uncorrelated with the level of discrimination earlier in the hiring process. The researchers conclude by discussing the substantive and methodological implications of our findings.

A previous version of this paper was presented at the April 2018 meetings of the Southern Sociological Society in New Orleans, LA. The authors received helpful comments from Michael Gaddis and Arnfinn Midtbøen. The authors gratefully acknowledge financial support for this project from the Russell Sage Foundation, Award #88-15-06.

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" rather than in a laboratory (National Research Council 2004; Gerber and Green 2012). And unlike reports of discrimination on 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 100 field experimental studies of hiring discrimination against minority races and ethnicities across more than 20 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. Instead, most studies stop with whether or not an applicant receives an invitation to an interview, often referred to as a "callback." Although researchers are interested in explaining racial disparities in hiring, 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 substantial amounts of employer time

[1] Authors' count; see Appendix B for more details. For other summaries of the literature on field experiments of discrimination see Baert (2018) and Zschirnt and Ruedin (2016).

(Cherry and Bendick 2018). As a result, only a small share of field experiments go to the job offer outcome; often these are studies backed by large grants or conducted with government support.

The widespread use of callbacks is understandable given the difficulties of conducting a field experiment that goes to the job offer outcome. Nevertheless, it leaves unanswered the following question: how adequate is our understanding of racial discrimination in the hiring process based on an assessment of racial differences in callback rates, when the ultimate subject of interest is discrimination in job offers? Does most of the racial discrimination in the hiring process actually occur when employers are deciding whom to invite for the interview? Or, even after majority and minority candidates make this first cut, are majority candidates still favored for reasons unrelated to their skills and other relevant qualifications? Finally, what does an understanding of the extent of discrimination across the various stages of the hiring process tell us about the nature of racial discrimination in hiring?

These questions should concern those interested in understanding the causes of persistent racial gaps in employment outcomes. 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 (Attström 2007; Pascoe and Richman 2009).

We address these questions by comparing callback and job offer outcomes in all field experimental studies of racial discrimination in hiring that go to the job offer. We use techniques from the meta-analysis literature, the branch of statistics concerned with combining results across studies. Our results indicate that substantial discrimination occurs even after minority candidates make it to the interview: about half of the discrimination in job offers results because of discrimination from application to callback.

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. The high causal (or internal) validity of field experiments for determining hiring discrimination results from the control that investigators exercise over applicants' characteristics and behaviors; this allows them to largely rule out differences other than race and ethnicity that might otherwise confound estimates of discrimination. Some field experiments are done in person, which we call in-person audits, while others have been conducted through the mail or over the internet, which we call resume audits.

In in-person audits, researchers send teams of trained actors to apply for the same job vacancies (Allasino et al. 2004; Cediey and Foroni 2008; Pager et al. 2009). Each team includes at least one actor who belongs to the native or dominant racial group and at least one actor from a racial minority background. Teams are assigned equivalent fictitious employment credentials like education, training, and previous experience. The majority and minority actors undergo a period of vetting and training that involves practice calls to employers, mock interviews, and standardizing candidate responses

(Bendick et al. 2010). Subsequently, actors are matched based on physical appearance, age, and demeanor, and then placed in teams. 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, physical appearance.[2]

In resume audit studies (also called correspondence studies), researchers submit resumes representing fictitious applicants by mail or over the internet. 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 common 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).

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 (perhaps saying they have accepted another job in the interim) or they simply do not respond.[3] A callback received signals a positive indication of employer interest and hence of success in the hiring process.

[2] See Appendix A for a list of the methods used in the 12 in-person audit studies included in our sample to signal the race or ethnicity of the minority applicants.

[3] In some studies stopping at callbacks, some job offer outcomes are observed because auditors are offered the job after applying and no interview is required (e.g., Pager, Western, and Bonikowski 2009).

However, a limited number of face-to-face audit studies have pursued applications all the way to the hiring 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.[5]

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. 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). Or, they may judge minority employees to be weaker prospective employees based on negative cultural stereotypes of minority group members (e.g., Bobo et al. 2012; Quillian and Pager 2010). Finally, employers may "statistically discriminate," or rely on on-average views of group members to make judgements about individual members of groups in the absence of detailed individual information (Arrow 1973). Some of the biases that affect hiring may even be unconscious, as demonstrated by studies of "implicit" attitudes (Greenwald et al. 1998; Rooth 2010).

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

[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.

[5] For a recent work on evidence of reverse discrimination in field experimental studies, see Bonoli and Fossati (Forthcoming).

below, there are reasons to believe that the level of discrimination may differ across stages. We divide the application process into two stages: application to callback and interview to job offer.

Perhaps the most compelling reason to expect more discrimination from application to callback than from interview to job offer is grounded in employer selection across the hiring process. The callback almost always 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 with a relatively low propensity to discriminate: employers with a high propensity to discriminate are likely to weed out identifiable minority applicants at the callback stage. Applicants who reach the interview will then tend to do so with employers who are less discriminatory; for this reason there may be less discrimination from interview to job offer than from application to callback.

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 posits that employers rely on group averages in making judgments about group members to the extent that they lack individual applicant information (Arrow 1973). The interview provides significant individual information about 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 is the basis for employer discrimination, then this additional information should result in employers relying less on group averages in making decisions, resulting in less discrimination from interview to job offer (Altonji and Pierret 2001).

Other theories, however, suggest the opposite pattern: discrimination at the final job offer stage might actually be quite significant. The interview is conducted face-to-face and tends to draw more interviewer attention. Some theories suggest these properties tend to increase discrimination.

First, there may be strong discrimination from interview to job offer because race is presented more clearly in the interview situation. Race-typed names provide good but not unambiguous signals about ethnicity (Gaddis 2017a, 2017b), and the clarity of these signals also depends on contextual factors like geographic location (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 from the name. Face-to-face appearances send a clearer signal that is less likely to be misunderstood, and may then produce more discrimination.

Second, in some cases the decision-makers in the interview and the job offer stages may simply be different. For instance, staff members in the human resources department may select interviewees but a supervisor may conduct 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 will then not operate as expected.

Third, racial stereotypes or race-related reactions 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 suggest that the general salience of race might be heightened in the context of face-to-face interactions (Greenwald et al. 1998). Rooth (2010) found that the probability of Arab applicants receiving a callback in Sweden was five percentage points lower among recruiters with more negative (or anti-minority) IAT scores.

Given the foregoing discussion, it is unclear how racial discrimination will play out across different stages of the hiring process. Discrimination during the callback stage is only an accurate indicator of the total level of discrimination if virtually all of the discrimination occurs during that initial stage and there is no “reverse” discrimination in favor of the minority group at the final stage. To assess how our view of discrimination is altered by considering the job offer, rather than 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 contrasts the prevalence of discrimination at three different stages of hiring: a pre-application inquiry as to whether the job is 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 declines across stages.

An important caveat to Zegers de Beijl's 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 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.

For instance, suppose 100 majority-minority auditor pairs initially apply for a job. At the voice inquiry stage, in 40 pairs both the majority and minority auditor are told the job is available, in 20 pairs the majority auditor is told the job is available and the minority auditor is told it is not, in 5 pairs the minority auditor is told the job is available and the majority auditor is told it is not, and in the remaining 35 pairs both auditors are told the job is no longer available. Zegers de Beijl estimates discrimination at the first stage as $\frac{20-5}{100} = 15\%$. In the second stage, only the 40 pairs of auditors that received equal treatment submit resumes. Suppose in 10 cases both are invited to interview, in 10 cases the majority auditor is asked to interview and the minority auditor is not, and in the remaining 20 cases neither auditor is invited to interview. Zegers de Beijl computes discrimination at the second 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 40 pairs of auditors submitted applications (made it to the second stage) and so were at risk of discrimination.

Zegers de Beijl measures discrimination as occurring whenever the majority applicant gets further in the hiring process than the minority applicant. By contrast, we focus on discrimination in receipt of job offers. We prefer to focus on discrimination in job offers because it is more closely linked to racial disparities in hiring and employment. 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

We perform a meta-analysis of all field experimental studies that go to the job offer outcome to examine how discrimination in callbacks compares to discrimination in job offers. 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. It is a standard method to combine results in fields that regularly conduct experiments (Borenstein et al. 2009).

In practice, our research design followed a three-step process. First, we identified all of the field experimental studies of racial discrimination in hiring that observe 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 using meta-analytic methods that combined the results of these studies, focusing on contrasting callback and job offer outcomes.

The search for studies and coding process were part of a larger project to gather and code information from all existing field experiments of racial and ethnic discrimination in hiring around the world. Details of how these procedures were conducted are discussed in Appendix B. More than 100 field experimental studies were gathered and coded. However, only thirteen studies go to the job offer– most field experiments stop at the callback.

Subsequently, we excluded one study, McIntosh and Smith (1974), because the callback sample was for skilled jobs, and the job offer sample was for unskilled jobs, using different tester pairs, thus making the callback and job offer outcomes somewhat incomparable. This study was also from the early 1970s, making it much older than the other studies in our sample. Sensitivity analysis that included McIntosh and Smith (1974) shows that our results are not substantively different if this study is included.

The twelve studies in our core sample follow respondents through two main stages: application to callback, and interview to job offer. In the first stage, testers submit a written application, resume, or other documents if necessary in response to job ads in the sample. In some audit studies the testers also called employers by telephone. The outcome of the first stage is receiving or not receiving a callback to be interviewed. In the second stage, testers who received a callback are actually interviewed, and the employers decide whether a job offer will be extended. The outcome at the second stage is the receipt of a job offer or not.

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

In the U.S.-based studies, racial minorities included Hispanics and blacks; across the European studies, racial minorities included candidates with backgrounds from South Asia, the Middle East, and North Africa, among others. 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 Appendix A for more details about the studies in our sample and how race is signaled at each stage.

To address whether our results based on a sample of 12 studies are generalizable to the broader population of field experiments of hiring discrimination, we contrast callback outcomes in our sample to callback outcomes in all other field experimental studies (i.e., those that stop at the callback) conducted in the eight countries for which we have job offer outcome data. This larger sample of studies with only callback outcomes includes 65 studies that encompass 96 estimates of discrimination against minority groups.

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 (interview invitations) 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. For example, if the discrimination ratio for callbacks equals 1.50, this implies that majority candidates received 50% more callbacks than equally qualified minority candidates. These are calculated from counts of outcomes available in each study report.

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 $\frac{c_w}{n_w} * \frac{n_m}{c_m}$.

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 $\frac{j_w}{n_w} * \frac{n_m}{j_m}$. Because in-person audit studies match groups on their non-racial characteristics either through the assignment of characteristics or through random assignment, no further within-study controls are required for valid estimates of discrimination. Discrimination ratios for the 12 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 16.0, indicating that native whites receive 36% fewer job offers to 16 times as many job offers as applicants from the minority group.[6] In only one

[6] The Bovenkerk1995a study that finds a discrimination ratio of 16.0 in job offers is something of an outlier among our studies. However, dropping this study has no effect on our main results, see footnote 9.

case does the minority group receive more job offers than the majority, Hispanics in the James and DelCastillo (1992) study.

Table 1 also shows the discrimination ratios for job offer conditional on callback. This is the discrimination ratio in job offers among respondents who successfully advanced to the callback stage. This conditional measure thus represents the additional discrimination that minority applicants face as they go from interview to job offer.

We focus on the discrimination ratio, but 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 (the ratio of proportions of callbacks/job offers, majority to minority), for reasons discussed in Appendix C. Appendix C also discusses results using these alternative measures.

To assess the difference in discrimination between callback and job offer, we log the callback and job offer discrimination ratios, and take the difference between each study's (logged) job offer and callback discrimination ratios. If y_{im}^{c} is the callback discrimination ratio for minority group m in the ith study, and y_{im}^{j} 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 $g_{im} = \ln\left(y_{im}^{j}\right) - \ln(y_{im}^{c})$. The ratios are logged to reduce the asymmetry of the ratio for analysis purposes, following standard practice in the meta-analysis literature. Values greater than 0 indicate more discrimination overall in receipt of job offers, and values less than 0 indicate more racial discrimination in the receipt of callbacks.

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 comparability measures is a more typical approach to do this in the meta-analysis literature.

The goal of a meta-analysis is to combine information across studies. The information each study provides is inversely proportional to the variance of the discrimination ratio. We calculate the variance of the ratio from counts reported in each study, accounting for audit pairs in the design when possible. To estimate this, we use standard formulas for the variability of a ratio due to sampling error and the counts of outcomes from each study. We also calculated the covariance of the discrimination ratio between callback and job offer stages for each study and use this to estimate the variance of the difference. Formulas are in Appendix D.

Meta-Analysis Statistical Model

Field experiments vary in their characteristics, such as the target group, geographic areas they cover, the exact job sectors covered, dates of the fieldwork, and the details of their methodology. To account for this variability we employ three procedures. First, we compare callbacks and job offers for the same studies. As discussed above, this implicitly holds constant similar variables in comparing callbacks and job offers. Second, to generalize the results beyond the 12 studies we observe, we employ a random effects specification (Raudenbush 2009). Random effects incorporate a variance component capturing variation in outcomes across studies that is due to unobserved study-level factors. Random effects are recommended whenever there is reason to believe that the effect in question is likely to vary as a result of study-level variables, rather than represent a single underlying effect that is constant over the whole population. 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 this study-level variability. Third, we use some models in meta-regressions with direct controls for some study characteristics, although the number of study characteristics we can control for is necessarily small because there are only 12 studies that go to the job offer outcome.

More formally, random-effects meta-analysis allows the true gap between callback and job offer outcomes in racial/ethnic discrimination to be estimated on average across studies by assuming that study gaps have a normal distribution around the population mean gap between callback 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 ($g_{im} = \ln\left(y_{im}^{j}\right) - \ln(y_{im}^{c})$), 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 from between-study variance as part of the meta-analysis model, while σ_{im}^2 is the variance of the logged response ratio of the mth minority group from whites in the ith study, estimated from study outcome counts as described above. Intuitively, the between study variance is estimated from the residual variation in study outcomes not accounted for by possible random sampling variation; in practice estimation is by restricted maximum likelihood (see Raudenbusch 2009). We also perform basic meta-analyses where the outcome is the callback or job offer logged discrimination ratio, $\ln(y^c)$ or $\ln(y^j)$, respectively.

Meta-regression adds predictors to this framework. It allows us to model the difference in discrimination ratios between job offer and callback as a function of a vector of k characteristics of the studies and effects, x, plus (in the random effects specification) residual study-level heterogeneity (between study variance not explained by the covariates). Principally, meta-regression allows us to investigate the association between discrimination at the callback stage and the job offer stage.

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

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

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). The estimation is by restricted maximum likelihood.

Small Sample Adjustments, and Accounting for Dependence of Discrimination Estimates

Our basic analysis is based on 12 studies. Samples of this size are common in meta-analysis (e.g. Brockwell and Gordon 2001). We employ small-sample corrections with clustered standard errors—discussed below—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 Hispanics (see Table 1 for a list of studies and groups). This is why we have 15 discrimination estimates ("effects") based on 12 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. Hedges et al. (2010) and Tipton (2015) discuss robust variance methods in meta-analysis. 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 (Fisher and Tipton 2015). [7]

Adjusting Discrimination Ratios in Studies with Conditional Following Rules

As a basic rule, we determine the outcomes from the point of initial application to the outcome of callback or job offer. Some studies report counts from initial application to callback and job offer directly, which we use to create these measures.

Five of our studies included a preliminary inquiry in which testers contacted employers by phone about job vacancies that were advertised and asked if the jobs were still available before applying. Employers would typically respond by agreeing to consider the application or by indicating that the vacancy had already been filled. This

[7] We also performed the same analyses using clustered robust standard errors with "hierarchical" weights, which is another method of accounting for dependence (Hedges et al. 2010; Tipton 2015). The results were substantively the same.

was not always true, because the other tester applicant sometimes received an invitation to submit a resume.[8] Applicants who were told a job was not available (and then did not submit a resume) were counted as not receiving a callback for the callback outcome.

A related complication is that some of these studies followed a conditional rule, in which a team of testers would only continue to the next stage if both were successful during the prior stage. Thus, if a minority candidate did not get an interview invite, the majority candidate would not attend the interview even if invited. This creates a missing data problem for applicants who did not attend an interview because their partner did not receive a callback.

We estimate success rates for these applicants to fill in these missing data. To do this, we assume that the average discrimination ratio for applicants whose partner did not receive an interview would be the same as the discrimination ratio among applicants in the same study who did receive an interview. We then estimate unconditional discrimination ratios by multiplying the discrimination ratio for getting a callback with the discrimination ratio for getting a job offer conditional on receiving a callback. Details of this procedure are in Appendix E.

Publication Bias

A potential problem in meta-analysis (and other reviews of literature) is publication bias, or the concern that studies with null effects might be less likely to be published, resulting

[8] In some of these cases, after hearing the foreign names or accents of the minority candidates during an initial phone inquiry, employers explicitly stated their unwillingness to hire and therefore interview the applicants (Allasino et al. 2004).

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. Significant correlations suggest possible publication bias (Sutton 2009).

We performed a typical test for publication bias for the job offer outcome with our sample, the Egger test. It failed to reject the null hypothesis of no significant correlation of effect size and sample size (p=.226). That is, there is no significant evidence of publication bias.

A lack of publication bias is not too surprising in the case of in-person audit studies that go to the job offer outcome, because the studies are sufficiently difficult and expensive to conduct that authors are likely to produce a publication or public report that is widely available 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 public reports be produced regardless of outcome. Moreover, given the existence of many studies that do document evidence of discrimination in hiring, null findings may also be viewed as sufficiently novel to support publication.

RESULTS

We begin with a basic meta-analysis of the level of discrimination at different stages. Our outcome is the discrimination ratio, which is a ratio of the success rate of the majority applicants to the success rate of comparable minority applicants. Results of the meta-analysis for each stage are shown in Figure 1 and Table 2 Panels A and B.

For our sample of 12 studies, the results indicate that majority applicants receive 52% more callbacks than equally qualified minority applicants (discrimination ratio of 1.523, 95% confidence interval of 1.32 to 1.76).

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 1.455; this indicates that even when both candidates receive an interview, majority applicants still receive about 46% more job offers than comparable minority applicants. Looking at the overall level of discrimination in job offers, majority applicants receive about 127.5% more job offers than comparable minority applicants (discrimination ratio of 2.275, 95% CI of 1.59–3.24). The difference between the callback discrimination ratio and the unconditional (or overall) job offer discrimination ratio is statistically significant at $p<.05$ (shown in Panel B of Table 2). On its face, 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 estimate suggests that there is more than twice as much discrimination overall in job offers as in callbacks.

Figure 2 shows a forest plot of the difference in the logged discrimination ratios for the callback and job offer outcomes for each of the 15 effect sizes (based on 12 studies) in our analysis. The study identifier and the minority group are displayed in the column on the left. The black square to the right of each study and minority group indicates the point estimate of the difference in the logged discrimination ratios for the callback and job offer outcomes, with the line providing a 95% confidence interval. In some cases, the confidence interval is very wide, especially for studies with small numbers of applications submitted or low success rates at each stage. The study weights, which provide a measure of the relative contribution of each study to the overall estimate, are also displayed in the column on the right. For 13 of the 15 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 for job offers than for callbacks indicate that the differences between the two stages are not statistically significantly (at p<.05). There is thus a fairly consistent pattern across studies of greater discrimination in job offers than in callbacks.[9]

How generalizable are our findings based on 12 studies that go to the job offer stage to the broader population of field experiments of racial discrimination in hiring? To help address this question, we compare the average callback discrimination ratio of the 12 studies in our sample that went to the job offer with the average callback discrimination ratio for all 65 studies located in our meta-analysis search (see Appendix B) that use the callback as the final outcome and were conducted in the same countries as the 12 studies. This includes 96 estimates of discrimination against minority groups.

Results are shown in Table 2 Panel C. The average (callback) discrimination ratio in the 65 studies that stopped at the callback is very similar to the average callback discrimination ratio in the 12 studies that went all the way to the job offer (1.547 vs. 1.523, respectively). This suggests that the levels of discrimination observed among the studies in our sample are not systematically different from those of the studies that stopped at the callback. In addition, there is a statistically significant difference between the discrimination ratio for the 65 studies that stop at the callback (1.547) and the job

[9] If we drop the Bovenkerk1995a study, which has a very high job offer discrimination ratio (16.0), this only decreases the meta-analysis estimated average job offer discrimination ratio from 2.275 to 2.240, leaving our conclusions unchanged. This study has a small effect on the average because it has a small sample size of applicants and receives a relatively low weight.

offer discrimination ratio (2.275). Overall, we would reach identical conclusions about the difference in levels of discrimination between the callback and job offer, regardless of whether we used the callback discrimination ratio from the studies that went all the way to the job offer or the ratio from those that did not.

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 are using 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, yet the strength of the association is unclear.

We find that high discrimination ratios at the callback stage are reasonably predictive of high discrimination ratios in the job offer stage. We estimate the correlation of the two stages to be about 0.57. Figure 3 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 unconditional job offer discrimination ratio tends to be greater than 0 at all levels of callback discrimination, but increases significantly with higher callback discrimination. This suggests that disadvantages for racial minorities tend to accrue across the stages of the hiring process.

Second, does the level of discrimination at the callback stage predict the magnitude of the additional discrimination that occurs at the final job offer stage? We find that the level of discrimination at the initial callback stage is uncorrelated with the extent of the additional discrimination that occurs during the final stage (i.e., conditional on having received an invitation to the interview). The line of best fit in predicting conditional job offer discrimination from callback discrimination is shown in Figure 4. It is almost flat.

The results suggest that employers who discriminate a great deal at the initial callback stage do not necessarily discriminate more against minority candidates who make it to the interview; conversely, employers who discriminate less at the initial stage might discriminate quite seriously against minorities at the final job offer stage. 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 (logged) discrimination ratios between callback and job offer is the outcome as a function of national, group, time, and author type covariates. Because there are only 12 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 USA vs. Europe, dummy variables for black/African and Middle-Eastern/North African target groups, a control for year of the study, and a control for whether or not the study was conducted by an advocacy group. Either alone or with other controls, none of the coefficients of these predictors are significant at conventional levels. The only regressor that is somewhat close to significance (p<.2) is black/African, which has a p-value of 0.15 in the model in which it is alone and a p-value of 0.14 with other predictors. 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: to detect discrimination. They suggest that auditors may then tend to act in accord with this purpose, resulting in auditors being more likely to find 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 predominance of resume audits (Cherry and Bendick 2018).

We are able to consider the Heckman-Siegelman critique of face-to-face audits by contrasting outcomes across studies done by different organizations.[10] Some audit studies have been conducted by advocacy groups to help document discrimination. Others are conducted by academics or government organizations. Given the strong interest of the advocacy groups in finding discrimination, there are reasons to believe that in-person audits run by advocacy groups should be especially likely to find high discrimination if auditors tend to "act the part" in a way that creates bias.

[10] 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.

If we assume that the desire to find discrimination is higher among persons in advocacy than other organizations, then we would expect larger disparities between callback and job offer outcomes for studies done by advocacy groups. According to our analysis in Table 3, this predictor is positive, consistent with the possibility of bias, although not statistically significant (in Models 4 and 5, p>.2). There is no clear evidence that the lack of a double blind design is a problem. The small sample size and imprecision, however, necessarily reduce the certainly of this conclusion. As such, it would be desirable if future research using in-person audits could be conducted double blind or could find ways to more directly address this potential complaint.

On the other hand, we also note that there are reasons to believe that in-person audits actually understate 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 natural ethnic presentation styles 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.

In cases where aspects of self-presentation are not relevant for job performance, field experiments may thus understate discrimination because they do not account for the discrimination grounded in cultural characteristics that are tied to a race or ethnic group. Rivera (2012) has recently argued that cultural matching between employers and workers plays an important role in hiring. This is underscored in 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). Yet in cases in which wearing a headscarf is not relevant to the job, this would constitute a form of ethnic discrimination. Audit studies may then underestimate discrimination by failing to capture a form of discrimination tied to ethnic-specific cultural norms and presentation styles that differ from those of the majority group.

DISCUSSION

Field experiments have become a basic tool for measuring the extent of racial discrimination in hiring. However, the vast majority of field experiments focus on racial disparities in callbacks, rather than on racial disparities in job offers. How does our view of racial discrimination in hiring change if we focus on job offers rather than on callbacks?

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 callback outcomes would suggest. Our point estimates indicate that there is more than twice as much discrimination against minorities in job offers as in callbacks, on average. This is a fairly consistent result in our data: in 13 of our 15 estimates of discrimination against minority groups, there is more discrimination in job offers than in callbacks. This results because 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). They found that on average, white applicants in the U.S. 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.[11]

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 level of total discrimination in hiring in job offers, 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 about half of the total discrimination in hiring occurs from initial application to callback, and because discrimination from interview to job offer is uncorrelated with the level of discrimination in callbacks (and thus does not tend to offset earlier discrimination, as it would if later discrimination was negatively correlated with earlier discrimination).

[11] 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%.

Our results also have implications for 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 meta-analysis also provides some evidence against the view that most racial discrimination in hiring reflects statistical discrimination. 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. Interviews increase this information because employers receive information in the interview regarding (at least) the dress, appearance, and demeanor of applicants. Contradicting this prediction, we find as much discrimination from interview to job offer as from application to callback.

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 12 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 12 studies we analyzed look very similar to that of the 65 known field experiments in these same countries that stopped at the callback, suggesting that the 12 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, it is difficult to ascertain the specific reason for this discrimination from our results. Assessing the specific reasons for the additional discrimination that occurs during the interview stage is an important topic for future research.

The vast majority of field experiments of racial discrimination in hiring use the callback as the outcome of interest. However, our results show that minority applicants face substantial additional discrimination even after they make it past the initial pile of resumes. Racial discrimination in the labor market is thus significantly more serious than is suggested by field experiments that stop with the callback. While 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 gender or religion. For discrimination on other bases, the correspondence between callback and job offer discrimination has yet to be established. In light of the high social and economic costs of discrimination in hiring, this is another potentially fruitful area for future research.

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.

Allasino, E., E. Reyneri, A. Venturini, and G. Zincone. 2004. *Labour Market Discrimination against Migrant Workers in Italy*. 67. Geneva, Switzerland: International Labour Organization.

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.” Pp. 3–33 in *Discrimination in Labor Markets,* edited by O. C. Ashenfelter and A. Rees. Princeton: Princeton University Press.

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

Baert, Stijn. 2018. “Hiring Discrimination: An Overview of (Almost) All Correspondence Experiments Since 2005.” Pp. 63–77 in *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*, *Methodos*, edited by S. M. Gaddis. 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, Alicia Simmons. 2012. “The Real Record on Racial Attitudes.” Pp. 38-83 in Social Trends in American Life, Peter V. Marsden, Ed. Princeton University Press.

Bonoli, Giuliano and Flavia Fossati. 2018. “More than Noise? Explaining Instances of Minority Preference in Correspondence Studies of Recruitment.” *Journal of Ethnic and Migration Studies* DOI: 10.1080/1369183X.2018.1502658

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*. 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.

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

Cherry, Frances and Marc Bendick. 2018. “Making it Count: Discrimination Auditing and the Activist Scholar Tradition.” Pp. 45-62 in *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*, edited by S. Michael Gaddis. Cham, Switzerland: Springer.

Crabtree, Charles and Volha Chykina. 2018. “Last Name Selection in Audit Studies.” *Sociological Science* 5:21–28.

Fisher, Zachary and Elizabeth Tipton. 2015. "robumeta: an R Package for Robust Variance Estimation in Meta-Analysis." **arXiv:1503.02220 [stat.ME]**

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, S. Michael, ed. 2018. *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*. Vol. 14. Springer.

Gerber, Alan S. and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. 1st ed. 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.

Heckman, James J. and Peter Siegelman. 1993. "The Urban Institute Audit Studies: Their Methods and Finding." Pp. 187–258 in *Clear and convincing evidence: Measurement of discrimination in America*, edited by M. Fix and R. J. Struy. Lanham, MD: Urban Institute Press.

Hedges, Larry V., Elizabeth Tipton, and Matthew C. Johnson. 2010. "Robust Variance Estimation in Meta - regression with Dependent Effect Size Estimates." *Research Synthesis Methods* 1(1):39–65.

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

McIntosh, Neil and David J. Smith. 1974. *The Extent of Racial Discrimination*. 547. London, UK: PEP.

National Research Council. 2004. *Measuring Racial Discrimination*. Panel on Methods for Assessing Discrimination. Rebecca M. Blank, Marilyn Dabady, and Constance F. Citro, Editors. Committee on National Statistics, Division of Behavior and Social Sciences and Education. Washington, D.C.: The National Academies Press.

Neumark, David. 2012. "Detecting Evidence of Discrimination in Audit and Correspondence Studies." *Journal of Human Resources* 1128-57.

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 Hana Shepherd. 2008. "The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets." *Annual Review of Sociology* 34:181–209.

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

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

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):10870–10875.

Raudenbush, S., 2009. “Analyzing Effects Sizes: Random Effects Coding,” in: Cooper, H.M., Hedges, L.V., Valentine, J.C. (Eds.), *The Handbook of Research Synthesis and Meta-Analysis*. Russell Sage Foundation, New York.

Rivera, Lauren. 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.” *Labor Economics* 17(3):523–34.

Sutton, A.J., 2009. Publication bias, in: Cooper, H.M., Hedges, L.V., Valentine, J.C. (Eds.), *The Handbook of Research Synthesis and Meta-Analysis*. Russell Sage Foundation, New York.

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. https://www.iza.org/publications/dp/10217/discrimination-against-female-migrants-wearing-headscarves

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

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.

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	Applications Submitted
Allasino2004	Italy	Moroccan	1.962	1.292	2.535	704
Arrijn1998	Belgium	Moroccan	1.871	1.279	2.394	1056
Attstrom2007	Sweden	Middle Eastern	2.046	0.923	1.888	854
Bendick1994	USA	African American	1.220	3.980	4.857	298
Bendick2010	USA	Tester Of Color	1.346	1.634	2.200	122
Bovenkerk1995A	Netherlands	Moroccan	1.704	9.391	16.000	230
Cediey2008	France	North African	1.670	2.760	4.609	1100
Cediey2008	France	Sub-Saharan African	1.870	4.100	7.669	620
Cross1990	USA	Hispanic	1.331	1.140	1.518	802
Hjarno2008	Denmark	Pakistani	1.512	1.102	1.667	206
Hjarno2008	Denmark	Turkish	1.364	2.017	2.750	260
James1992	USA	Black	1.031	1.103	1.138	290
James1992	USA	Hispanic	1.071	0.597	0.640	280
Prada1996	Spain	Moroccan	1.689	1.917	3.237	528
Turner1991	USA	Black	1.144	1.306	1.494	992

Note: 12 studies, 15 effects (discrimination estimates against a distinct target group)

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

A. Stage of Hiring (N = 12 studies, 15 effects)	**Mean Discrimination Ratio**		**95% CI**		**Tau-squared**
Callback	1.523	***	1.318	1.761	0.0532
Job Offer conditional on Callback	1.455	*	1.079	1.961	0.0972
Job Offer Unconditional	2.275	***	1.592	3.252	0.1468

B. Differences Between Stages (N = 12 studies, 15 effects)	**Mean Difference in Log Discrimination Ratios**		**95% CI**		**Tau-squared**
Callback and Job Offer Conditional on Callback	-0.032		-0.373	0.309	0.1603
Callback and Job Offer, Unconditional	0.368	*	0.074	0.661	0.0863

C. Comparison to Field Experiments with Callback Outcome Only (65 Studies, 96 Effects)	**Mean Discrimination Ratio**		**95% CI**		**Tau-squared**
Callback, Field Experiments with Callback Outcome Only	1.547	***	1.414	1.693	0.0768
Log Difference Callback Outcome (65 studies) and Job Offer Unconditional (12 studies)	0.369	*	0.017	0.720	

* = p<.05; ** = p<.01; *** = p<.001

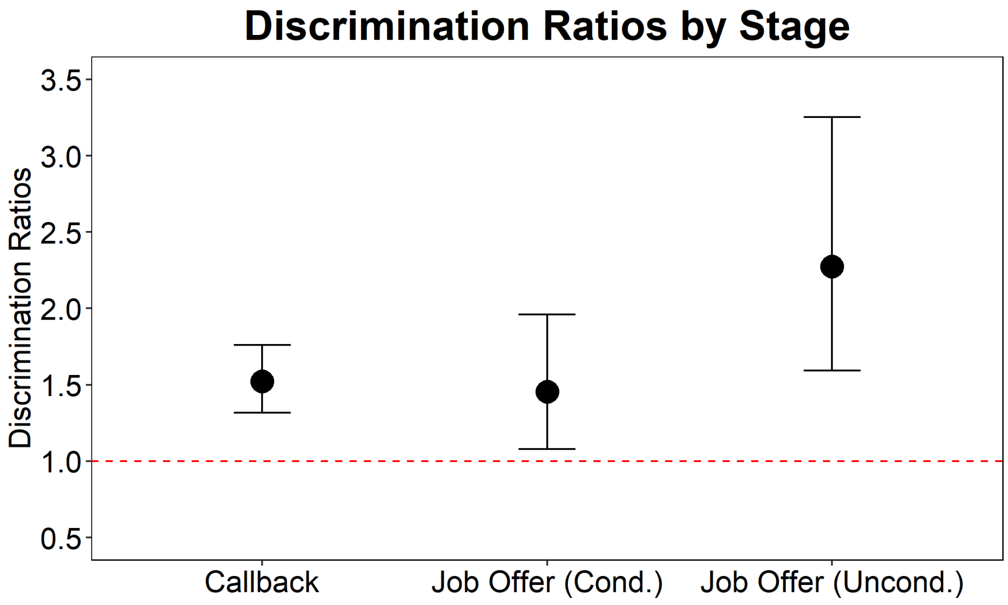
Note: Random-effects meta-anlaysis with clustered standard errors at the study level, using "correlated" cluster weights (Tipton 2015) with assumed rho=.8. Mean discrimionation ratios in panels A and C are based on the anti-log of the mean logged discrimination ratio. Tau-squared is the estimated between-study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

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

Outcome: Difference of Log Job Offer and Callback Discrimination Ratios	(1)	(2)	(3)	(4)	(5)
Country = USA (vs. Europe)	-0.14				-0.62
	(0.27)				(0.42)
Minority Group Black/African (1=yes)		0.61			0.57
		(0.35)			(0.31)
Minority Group Middle-Eastern/North African (1=yes)		0.30			-0.33
		(0.21)			(0.46)
Year of fieldwork of study (four digit year)			0.01		0.00
			(0.02)		(0.02)
Audit conducted by Advocacy Organization (1=yes)				0.32	0.39
				(0.27)	(0.28)
Constant	0.44	0.08	-19.52	0.27	-0.39
	(0.19)	(0.15)	(49.65)	(0.16)	(43.28)
Tau-squared	0.1107	0.1452	0.1136	0.0888	0.1454
N Studies / N Effects	12/15	12/15	12/15	12/15	12/15

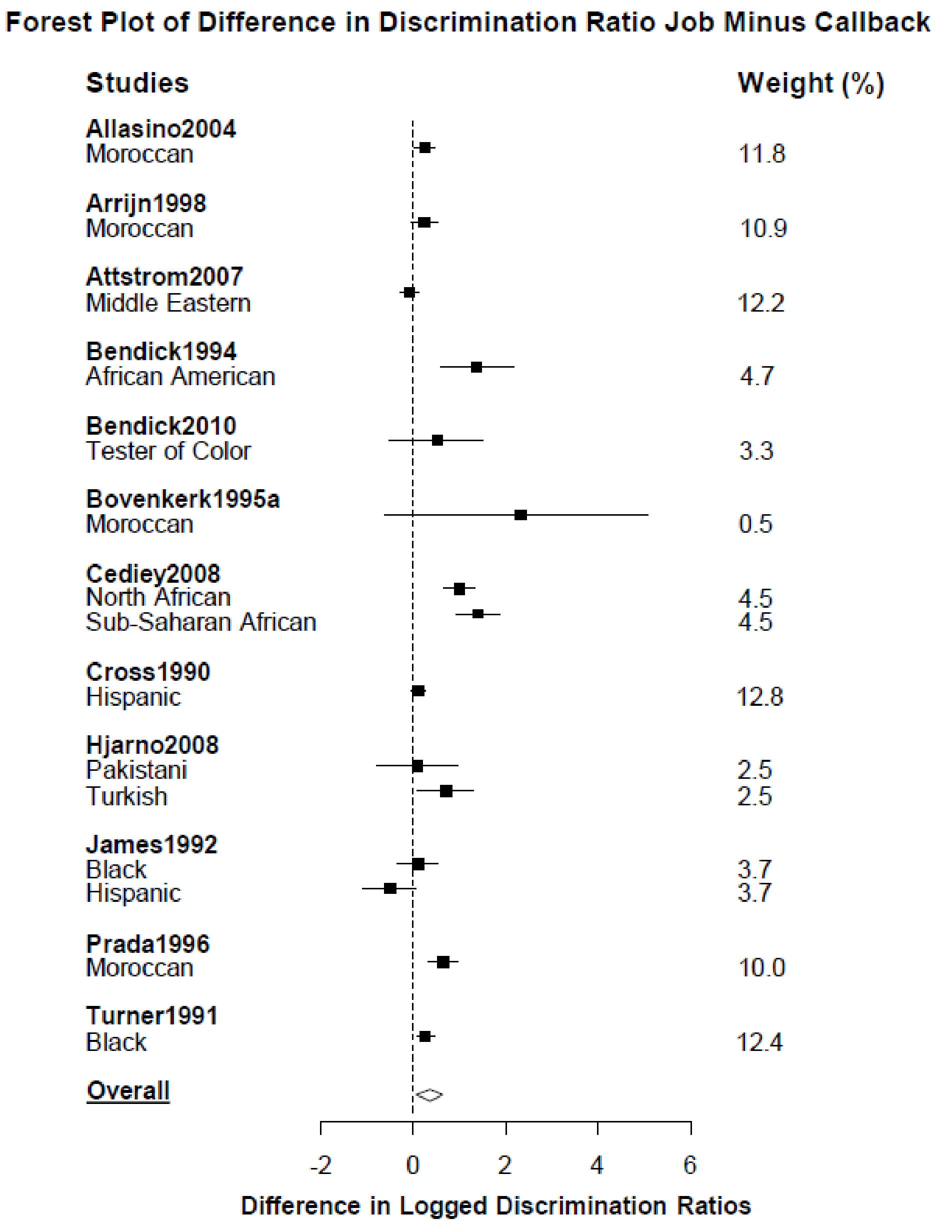
Note: Random-effects meta-regression with clustered standard errors at the study level, using "correlated" cluster weights (Tipton 2015) with assumed rho=.8.

Figure 1:



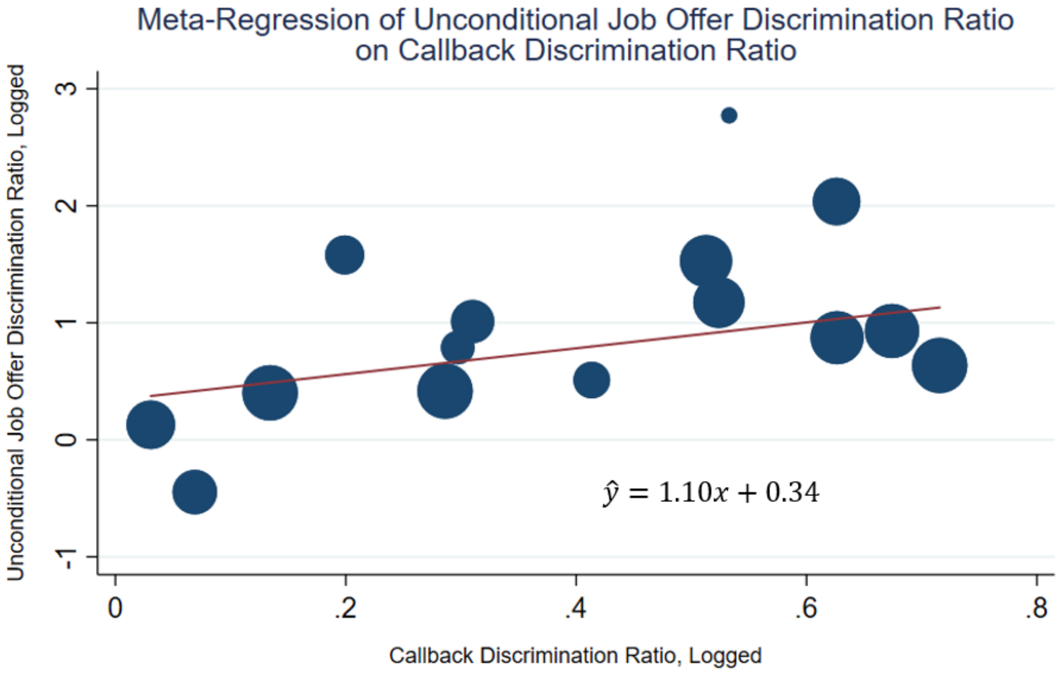
Note: Dots indicate point estimates of mean discrimination ratio from meta-analysis shown in Table 2 Panel A. Lines indicate 95% confidence intervals. Dotted red line is discrimination ratio of 1.0 (no discrimination).

Figure 2:



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

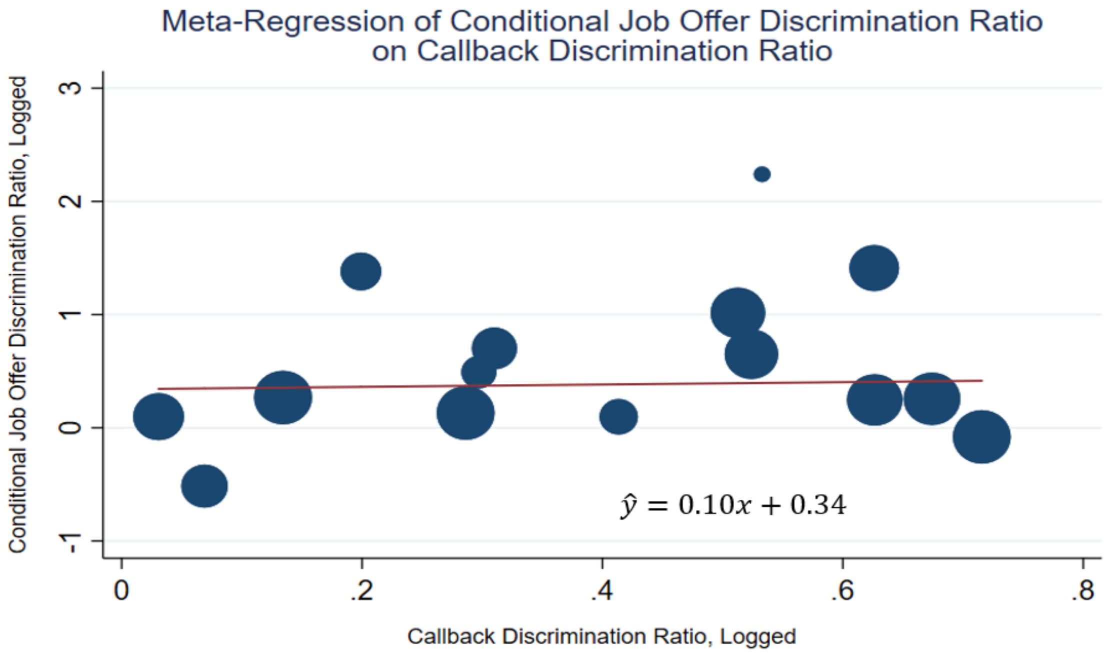
Figure 3:



Meta-Regression of Unconditional Job Offer Discrimination Ratio on Callback Discrimination Ratio

Note: Meta-regression with robust standard errors. Marker sizes are proportional to the weights used in the model.

Figure 4:



Meta-Regression of Conditional Job Offer Discrimination Ratio on Callback Discrimination Ratio

Note: Meta-regression with robust standard errors. Marker sizes are proportional to the weights used in the model.

Appendix A: Testing Characteristics of Studies going to the Job Offer Outcome

Study ID	Target Groups	How the Testers were Matched	Application Stages (what occurs, outcome)	Minority Status Signal
Allasino2004	Moroccan	Matched on: personal characteristics (e.g., qualifications, skills), and demeanor (p. 36). The Moroccan candidate had an accent (p. 37, 43, 47, 57); the minority group in this study was comprised of first-generation immigrants from Morocco.	[Stage 1]: voice inquiry by phone, allowed to apply; [Stage 2]: written application, interview invite; [Stage 3]: interview, job offer (p. 39-40)	[Stage 1]: Arab accent, Moroccan name (p. 43); [Stage 2]: name; [Stage 3]: accent, name, and physical appearance (p. 48)
Arijn1998	Moroccan	The testers were matched on age, physical characteristics ("weight, height, build"), language skills, demeanor, behavior (p. 28-29). "Candidates of Moroccan origin should not speak with a foreign accent" (p. 29). Candidates of Moroccan origin communicated their race via name (e.g., Ahmed, Fouad) (p. 49, 56, 73).	[Stage 1]: voice inquiry by phone, written resume, or direct presentation; [Stage 2]: interview; longer phone call; or submission of CV2; [Stage 3]: job offer or invitation to trial period (p. 27-28)	[Stage 1]: name (p. 63); [Stage 2]: name; [Stage 3]: name
Attstrom2007	Middle Eastern	No explicit mention of a foreign accent. The minority group consisted of native-born Swedes of Middle Eastern origin (p. 22). The minority applicants were fluent in Swedish. Matching characteristics: "actual appearance, body language, attitude, personality" (p. 2).	[Stage 1]: voice inquiry by phone, allowed to apply; [Stage 2]: written application, interview invite; [Stage 3]: interview, job offer (p. 18-19)	[Stages 1/2]: Arab name (ps. 30, 32-33); [Stage 3]: name and physical appearance
Bendick1994	African American	"Each pair teamed one African American research assistant with a white research assistant of the same sex, approximate age, personal appearance, articulateness, and manner" (p. 27). The testers were also assigned a similar set of educational credentials, work experiences, and job-relevant skills (p. 28).	[Stage 1]: in person application, interview invite [Stage 2/3]: interview, job offer (p. 29-31, 40)	[Stage 1]: skin color; [Stage 2/3]: skin color (p. 27)

Appendix A, Continued: Testing Characteristics of Studies going to the Job Offer Outcome

Study ID	Target Groups	How the Testers were Matched	Application Stages (what occurs, outcome)	Minority Status Signal
Bendick2010	Tester of Color	Matched in relation to "gender and...age, appearance and manner" (p. 808). However, some of the majority and minority testers had a "slight accent" (ps. 808, 814). For example, one of the majority testers, a white woman, had a slight French accent (p. 810).	[Stage 1]: application, interview invite; [Stage 2]: interview, job offer (p. 808-809)	[Stage 1]: name; [Stage 2]: accent (sometimes), name, physical appearance
Bovenkerk1995a	Moroccan	Matched in relation to "...conventional appearance (average weight, average height, conventional dress and hair, and conventional dialect)" and "overall demeanour, openness, enthusiasm and communicative abilities" (p. 4). The Morrocan testers were born in the Netherlands, so it is assumed that they did not have a foreign accent (the study did not explicitly clarify).	[Stage 1]: voice inquiry by phone, allowed to apply; [Stage 2]: interview invite; [Stage 3]: interview, job offer (p. 8-9). Note: the distinction between the 1st and 2nd stages is a bit unclear in this study.	[Stage 1]: Moroccan name (p. 8); [Stage 2]: name; [Stage 3]: name and physical appearance (p. 9)
Cediey2008	North African, Sub-Saharan African	Matched per age (20-25) and "appearance – standard and similar clothing, level of expression" (p. 106)	[Stage 1]: voice inquiry by phone or resume submission, interview invite; [Stage 2/3]: interview, job offer (p. 59-61)	[Stage 1]: name; [Stage 2/3]: name, physical appearance
Cross1990	Hispanic	There were "16 testers with equivalent personal information, work and educational histories, and references" (p. 23). The testers were also trained to answer interview questions "with similar levels of enthusiasm, length of responses, and demeanor." The main difference was that the Hispanic testers had slight Spanish accents and light brown skin.	[Stage 1]: voice inquiry by phone, allowed to apply; [Stage 2]: written application, interview invite; [Stage 3]: interview, job offer (p. 40-41)	[Stage 1]: name, accent; [Stage 2]: name; [Stage 3]: name, accent, and physical appearance

Appendix A, Continued: Testing Characteristics of Studies going to the Job Offer Outcome

Study ID	Target Groups	How the Testers were Matched	Application Stages (what occurs, outcome)	Minority Status Signal
Hjarno2008	Pakistani, Turkish	The minority applicants did not have a foreign accent (p. 3, 25). Matching: appearance, accent, language ability (p. 8).	[Stage 1]: voice inquiry by phone, allowed to apply; [Stage 2]: written application, interview invite; [Stage 3]: interview, job offer (p. 11-12)	[Stage 1]: name; [Stage 2]: name; [Stage 3]: name and physical appearance
James1992	Hispanic, Black	Two-person teams of young men were employed; each had one minority and one white auditor. Both were matched on personal appearance, dress, personality, education, background, and other characteristics (p. 38, 40). Hispanic auditors were matched on nationality (native born), and they did not have accents. However, they all had distinctly "Hispanic looking" skin and hair color (p. 39).	[Stage 1]: personal visit to employer (required for the black-white pair) or voice inquiry by phone, allowed to apply or offered a job on the spot [Stage 2]: written application, interview invitation or contact info taken; [Stage 3]: interview, job offer (p. 42)	Hispanic: [Stage 1]: name or appearance (if in person visit); [Stage 2]: name; [Stage 3]: name and appearance (p. 39); Black: [Stage 1]: appearance; [Stage 2]: no signal for the written application [Stage 3]: appearance
Prada1996	Moroccan	Matching based on physical appearance (e.g., weight, height, dress), qualifications, and demeanor (p. 24). The Moroccan candidates communicated their national/ethnic identity through their distinctive accent, name, statement of nationality (p. 30)	[Stage 1]: phone call to express interest in job, interview invite; [Stage 2]: interview; [Stage 3]: job offer (p. 29-31)	[Stage 1]: name, accent, and statement of nationality [Stage 2/3]: name and accent (p. 45)
Turner1991	Black	"Conventional appearance was the major selection criterion--average height, average weight, conventional dialect, and conventional dress and hair. This made audit partners potentially interchangeable..." (p. 25).	[Stage 1]: in person application submission; interview invite [Stage 2]: interview [Stage 3]: job offer (p. 31)	[Stage 1]: appearance; [Stage 2] appearance [Stage 3]: appearance (p. 25)

APPENDICES (TEXT)

Some of the text below reprises discussions from the text and supplemental materials to Quillian et al. (2017).

Appendix B: Study Search Methods

Our search for field experiments started with a bibliographic search. Our search covered the following bibliographic databases and working paper repositories: Thomson's Web of Science (Social Science Citation Index), ProQuest Sociological Abstracts, ProQuest Dissertations and Theses, Lexis Nexis, Google Scholar, and NBER working papers. We searched for some combination of "field experiment" or "audit study" or "correspondence study" and sometimes included the term "discrimination", with some variation depending on the search functions of the database. To improve our coverage of non-English publications, we also searched two French-language indexes, Cairn.info and Persée, and two international sources, IZA discussion papers, a German working paper archive, and ILO International Migration Papers. Finally, we conducted a search with Italian, Spanish, Portuguese and Dutch translations of the search terms and other terms frequently used in these languages to describe field experiments in hiring discrimination in Google Scholar. The search was first performed in March 2014 and repeated in August and September 2014 and in November 2015. Searches in Italian, Spanish, Portuguese, and Dutch were conducted in November 2015 and February 2015. Our main search ended in 2015. We added few new studies after that time, except for updated versions of studies we had already located by December 2015.

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

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

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

Overall our search located more than 100 studies and included contrasts between white and non-white groups who were on-average equivalent in their labor-market relevant characteristics (e.g. education, experience level in the labor market, etc.), and who otherwise met our inclusion criterion.[1] Some of these studies included contrasts between more than one target group and whites (e.g. blacks and Hispanics) producing multiple estimates of discrimination against non-whites. However, only 14 studies in eight countries had auditors pursue job opportunities all the way to the job offer outcome. Another 65 studies in the same eight countries stopped at the callback outcome, which we use as the outcome in our study.

Appendix C: Alternative Measures of Discrimination

Two other measures that could be used instead of the discrimination ratio are the difference in proportions of positive responses or the odds ratio.

The ratio has the advantage in interpretation that it directly generalizes to indicate the relative number of applications that would need to be submitted by majority and minority applicants to be expected to receive the same number of job offers. For instance a ratio of 1.50 indicates that a minority applicant would need to submit 15 applications for each 10 by majority applicants to expect to receive the same number of positive responses.

By contrast, to understand the difference in proportion it is necessary to invoke the overall rate of positive responses. For the difference in proportions, high base rate studies dominate low base rate studies in terms of the measure: for instance, a study where 44% of whites and 40% of blacks receive a positive response gives the same discrimination difference estimate as one where 8% of whites and 4% of blacks receive positive responses, although our view is the latter shows much higher discrimination than the former. In general we prefer the discrimination ratio to the difference in proportions because it is less sensitive to the base rate of the outcome (see Bornstein, Hedges, Higgins, Rothstein 2009, chapter 5).

Another potential choice with good statistical properties is the odds ratio rather than the ratio of proportions, but we prefer the ratio of positive responses because it is much more easily interpretable. We also estimated our basic results using the odds ratio outcome, which produced similar results to those we find with the discrimination ratio.

Appendix D: Meta-Analysis Effect Weighting, Variance, and Covariance

[1] We excluded some studies where it was not clear if employers were the ones making decisions producing discrepant outcomes because applications were conducted through an employment agency. A few other studies were excluded because they lacked basic information on counts of outcomes by target group needed to conduct our analysis and the authors could not be located or declined to provide this data when contacted.

We calculate the standard error of the ratio from counts reported in each study, accounting for audit pairs in the design when possible. To estimate this we use standard formulas for variability of a ratio due to sampling error and the counts of outcomes from each studies. 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:

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

This is from Bronstein, Hedges, Higgins, and Rothstein (2009, formula 5.3). For studies that use and report a paired design – with one minority and one white applicant applying for each job – we use an alternative formula to account for the pairing (Zou 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, and p^d be the number of pairs in which neither tester received a callback. The variance of the logged odds ratio for the mth minority group in the ith study with paired data is:

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

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

For studies that are not paired between whites or nonwhites or where paired outcomes are not reported, we use formulas for the standard error for unpaired groups. This formula will slightly over-estimate the standard error of the effect for studies that are paired but we treat as unpaired due to lack of information about the outcomes at the pair level, underweighting these studies a bit in computing the overall effect, and slightly inflating the overalls cross-study standard error.

For some analyses we use the difference in the log job offer discrimination ratio and the log callback discrimination ratio, g_{im} ($g_{im} = \ln\left(y^j_{im}\right) - \ln(y^c_{im})$). There is positive covariance (correlation) at the study level between the job offer and callback discrimination ratio. We calculate the correlation in the callback rates job offer to callback using the fact that the probability of a job offer is zero when a callback does not occur, and assuming no association of callback and job offer discrimination ratios conditional on receipt of a callback. The covariance of the callback(c) and job offer(j) outcomes for each effect size (in each study) can be calculated as:

$$cov(c,j) = \frac{n_{c=0}\bar{c}\bar{j} + n_{j=1}(1-\bar{c})(1-\bar{j}) + n_{c=1,j=0}(1-\bar{c})(-\bar{j})}{n}$$

And the correlation of callback and job offer is:

$$cor(c,j) = \frac{cov(c,j)}{\sqrt{\bar{c}(1-\bar{c})}\sqrt{\bar{j}(1-\bar{j})}}$$

Where $\bar{c}$ is the proportion of callbacks received, $\bar{J}$ is the proportion of job offers (of all applications) received, n is the total number of applications submitted for the majority and minority group, $n_{c=0}$ is the number of applications for which no callback was received, $n_{j=1}$ is the number of applications for which a job offer was received, and $n_{c=1,j=0}$ is the number of applications for which a callback was received but no job offer was received. Finally we calculate the covariance between the log risk ratio of job offer and callback using the estimated correlation and formula 1.13 from page 1198 of Wei and Higgins (2012). This covariance is then used in calculating the variance of the difference between the log job offer and log callback discrimination ratios.

Appendix E: Adjustment to Discrimination Ratios in Studies with a Pre-Application Stage

A few studies in our sample follow a multi-stage design in measuring discrimination. This was a study design used by some European studies commissioned by the International Labor Organization. In these studies the applicants first called employers by phone to inquire if a job was still available. We would like to incorporate these responses into our measures of discrimination, to get total discrimination from initial application to the callback. For situations where either both applicants were told the job is still available or told it is not available, this is straightforward because we know if the callback or job offer is ultimately received.

In five studies, if one applicant was told the job was available and the other was not, no application was submitted by *either* tester. This last aspect of this design – that when one applicant received a positive response and the other did not, the applicant who could have then submitted a resume did not – requires some adjustment. We want to capture rates of receiving a callback (or job offer) for all minority and majority applicants from the point of initial application. We know that respondents who were told “no job is available” did not receive a callback (or job offer). But when one member of a pair was told the job is available, and the other was not, we do not know how often the member of the pair who was told the job was available would have received a callback (or job offer) if they had applied. We need to estimate this to get complete callback (or job offer) outcomes from the point of application.

To estimate callback and job offer rates in these studies, we assume that the member of the pair who received the invitation to interview but did not submit a resume (because their partner was told the job was no longer available) was as likely to get a callback or a job offer if they had submitted a resume as applicants of the same race/ethnic group in the same study for which an application was submitted.

More formally, we adjust the discrimination ratios in these five studies in the following way. Define:

n_1^w is the number of applicants from the native majority (white) group who initially call the employer to inquire if jobs are available. b_1^m is the number of applicants from the minority group who initially call the employer to inquire if jobs are available.

f_1^w is the number of applicants from the native majority (white) group who are told the job is still available. f_2^m is the number of applicants from the minority group who are told the job is still available.

n_2^w is the number of applicants from the native majority (white) group who submit application materials. n_2^m is the number of applicants from the minority group who submit application materials.

c_2^w is the number of applicants from the native majority (white) group who receive a callback. c_2^m is the number of applicants from the minority group who actually receive a callback.

We calculate the estimated discrimination ratio for minority group j in study i from the point of initial application with:

$$Y_{ij}* = \left(\frac{f_1^w / n_1^w}{f_1^m / n_1^m}\right)\left(\frac{c_2^w / n_2^w}{c_2^m / n_2^m}\right)$$

This just multiplied the discrimination ratio at the stage of asking if job is still available with discrimination ratio at the stage of receiving a callback. We use this estimated discrimination ratio for these five studies.

We calculate the estimated variance of the log adjusted discrimination ratio with:

$$\mathrm{Var}(\ln(Y_{ij}*))=\frac{1}{n_1^w\left(f_1^w / n_1^w\right)\left(c_2^w / n_2^w\right)}-\frac{1}{n_1^w}+\frac{1}{n_1^m\left(f_1^m / n_1^m\right)\left(c_2^m / n_2^m\right)}-\frac{1}{n_1^m}$$

This is the standard formula for the variance of a risk ratio with unpaired groups (Bornstein, Hedges, Higgins, and Rothstein 2009, formula 5.3), substituting the implied count of successes based on our estimation. Using the unpaired formula for these paired studies will somewhat overstate the variance of the ratio, while treating the counts as actual rates than estimated somewhat understates it.

REFERENCES (Appendix)

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

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.

Wei, Yinghui and Julian PT Higgins. 2013. “Estimating within-Study Covariances in Multivariate Meta-Analysis with Multiple Outcomes.” *Statistics in Medicine* 32(7):1191–1205.

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.