Journal of Economic Literature 2018, 56(3), 799–866
https://doi.org/10.1257/jel.20161309

Experimental Research on Labor Market Discrimination†

David Neumark*

Understanding whether labor market discrimination explains inferior labor market outcomes for many groups has drawn the attention of labor economists for decades—at least since the publication of Gary Becker's The Economics of Discrimination *in 1957. The decades of research on discrimination in labor markets began with a regression-based "decomposition" approach, asking whether raw wage or earnings differences between groups—which might constitute prima facie evidence of discrimination—were in fact attributable to other productivity-related factors. Subsequent research—responding in large part to limitations of the regression-based approach—moved on to other approaches, such as using firm-level data to estimate both marginal productivity and wage differentials. In recent years, however, there has been substantial growth in experimental research on labor market discrimination—although the earliest experiments were done decades ago. Some experimental research on labor market discrimination takes place in the lab. But far more of it is done in the field, which makes this particular area of experimental research unique relative to the explosion of experimental economic research more generally. This paper surveys the full range of experimental literature on labor market discrimination, places it in the context of the broader research literature on labor market discrimination, discusses the experimental literature from many different perspectives (empirical, theoretical, and policy), and reviews both what this literature has taught us thus far, and what remains to be done.* (*JEL* C93, J14, J15, J16, J23, J41, J71)

1. *Introduction*

Understanding whether labor market discrimination explains inferior labor market outcomes for many groups has drawn the attention of labor economists for decades—at least since the publication of Gary Becker's *The Economics of Discrimination* in 1957. The question is of obvious importance for public policy, as the answer can help policy makers determine whether (and perhaps how much) to emphasize efforts to combat discrimination in trying to bring about greater equality between racial and ethnic groups and between men and women, and in trying to improve the circumstances of other groups that suffer economic disadvantages, such as the disabled.

* University of California, Irvine, NBER, and IZA. I am grateful to Stijn Baert, Marc Bendick, Lieselotte Blommaert, Steven Durlauf, Peter Kuhn, Matt Notowidigdo, Dan-Olof Rooth, Devah Pager, David Phillips, Patrick Button, Bradley Ruffle, Doris Weichselbaumer, Uri Zak, and anonymous referees for helpful comments.

† Go to https://doi.org/10.1257/jel.20161309 to visit the article page and view author disclosure statement(s).

The decades of research on discrimination in labor markets began with a regression-based "decomposition" approach, asking whether raw wage or earnings differences between groups—which might constitute prima facie evidence of discrimination—were in fact attributable to other productivity-related factors. Subsequent research—responding in large part to limitations of the regression-based approach—moved on to other approaches, such as using firm-level data to estimate both marginal productivity and wage differentials.

In recent years, however, there has been substantial growth in experimental research on labor market discrimination—although the earliest experiments were done decades ago. The growth in the experimental research on discrimination likely in part reflects the growth of experimental research in economics generally. However, it is also a specific response to continuing challenges in drawing definitive conclusions from nonexperimental research about the role of labor market discrimination. Some experimental research on labor market discrimination takes place in the lab. But far more of it is done in the field, which makes this particular area of experimental research unique relative to the explosion of experimental economic research more generally.

My goal in this survey is to cover the full range of experimental literature on labor market discrimination, to place it in the context of the broader research literature on labor market discrimination, to discuss the experimental literature from many different perspectives (empirical, theoretical, and policy), and to review what this literature has taught us thus far about the existence of discrimination, which groups it affects, and many other questions—as well as what remains to be done. Moreover, one point of emphasis is summarizing and evaluating the evidence from many recent experimental studies (and some earlier ones) that try to determine the nature of discrimination—whether taste-based, statistical, or implicit. The distinction is important for policy, and recent research has been creative in trying to augment experimental studies with tests for the nature of discrimination. I suggest that the tests to this point are not very convincing, but also discuss how the studies implementing them have pointed the way to how researchers might more convincingly conduct such tests.

There are points of overlap with other recent research reviews. Anderson, Fryer, and Holt (2006) provide a brief review of laboratory experiments on discrimination covering a small set of studies on race or ethnicity, excluding studies that ask subjects to make decisions about fictitious workers. Pager (2007) provides an overview of a small number of field experiments on race discrimination, focusing on differences between the clear evidence of race discrimination from these studies and the suggestion from some analyses of secondary data that race no longer matters much in labor markets. Riach and Rich (2002) and Rich (2014) provide summaries of many field experiments on discrimination in labor markets; the thoroughness of these reviews enables me to avoid a detailed cataloging of results from many studies, and to instead discuss experimental research on labor market discrimination in the context of the broader literature and debates about discrimination, to focus to some extent on methods, and to delve into what I view as the most important issues that these studies confront or still have to confront.[1] In addition,

[1] Two reviews published in this journal are further afield, yet related. Harrison and List (2004) provide a general review of field experiments, focused largely on defining a typology of experiments, highlighting what they can teach us in comparison to other kinds of studies, and illustrating the kinds of questions they can answer. There is little overlap, although one section of their paper has a few references to audit or correspondence studies of discrimination. Lang and Lehmann (2012) provide what is largely a survey of theoretical models of race discrimination, with some

there has been a large number of experimental studies done in the last few years that figure prominently in some parts of this review. Clearly, these other reviews can be usefully read along with this survey to get an even fuller overview and understanding of what economists have learned and still might learn from experimental research on labor market discrimination.

2. *The Facts We Are Trying to Understand*

The raw data that motivate most of the research on discrimination are generally in the direction consistent with discrimination—e.g., lower pay for groups thought to experience discrimination in the labor market. But untangling whether discrimination in fact generates these differences is difficult, which explains the explosion of experimental research on labor market discrimination. Here, I briefly describe many of the stylized facts that underlie the discrimination literature, drawing on other work. I defer, until section 4, discussion of nonexperimental work testing discriminatory versus nondiscriminatory explanations of these differences, except to touch on some simple partial correlations, when available.

The experimental research focuses on hiring, and thus might be thought of as most relevant for understanding group differences in employment and unemployment. However, the segregation evidence for women discussed below shows that hiring discrimination may also lead to lower earnings. Moreover, models show that discrimination in hiring can lead directly to lower earnings for groups that experience this discrimination.

2.1 *Race, Ethnicity, and Sex*

In the United States, there has been a persistent raw difference between the earnings of black and white men—a bit over 20 percent over the past decade, and a shade larger for full-time, year-round workers (Lang and Lehmann 2012). Some factors explain part of this gap; in particular, Neal and Johnson (1996) showed that controlling for age and performance on the Armed Forces Qualifying Test (AFQT) diminishes the race gap for young men substantially. But other factors cut in the other direction, such as schooling (conditional on the AFQT, blacks get more schooling). Lang and Lehmann also note that earnings differences between black and white women have historically been lower and sometimes even reversed. Wages of Hispanics also lag substantially behind those of whites, with about 32 percent lower median weekly earnings for full-time workers (Kochhar 2008). The gap is about twice as big for foreign-born as native-born Hispanics.

Pay differences between men and women in the United States have fallen since the beginning of the 1980s, but still persist, with the gap in annual earnings for full-time, year-round workers recently narrowing to around 22 percent, and 18 percent for weekly earnings (Blau and Kahn 2017). Some samples now show women getting more education, but men have tended to study more lucrative fields. Lower work experience of women also plays a role, but this difference has declined as well.

Employment (and unemployment) gaps between blacks and whites are pronounced in US data. Lang and Lehmann (2012) report a difference in the labor force participation rate of black versus white men aged 25–54 of 7.8 percentage points in 2008, and an unemployment rate difference of 4.6 percentage points. The unemployment rate differential appears unrelated to education, and there is a large earnings differential (which, in contrast

emphasis on drawing empirical implications. Again, the overlap is minor, although their discussion connects with the material I discuss on distinguishing between taste and statistical discrimination. In addition, there is a brief discussion of audit/correspondence studies that touches on some of the points I cover in detail.

to wages, also reflects hours and employment differences that persist when controlling for AFQT scores (Johnson and Neal 1998)). Hispanics have participation rates that slightly exceed those of whites, but unemployment rates are nearly a third higher (Kochhar 2008). Women continue to have lower employment rates than men; their labor force participation rate in 2013 was lower by 12.5 percentage points. However, their unemployment rates are slightly lower than men's.[2]

There is similar prima facie evidence of discrimination in other countries for groups that are the focus of experimental work on discrimination. For example, there are lower employment rates for indigenous people and immigrant groups in Australia, although these become smaller and insignificant with controls for language proficiency and schooling (Booth, Leigh, and Varganova 2012), and many studies of Europe (including some in the references, such as Bartolucci 2014, and Blommaert, Coenders, and van Tubergen 2014) document lower earnings and employment for minorities or immigrants.

Discrimination in hiring may affect where a person works with, for example, important implications for earnings if women are hired into lower-paying occupations or firms (e.g., Groshen 1991).[3] There is also substantial workplace segregation by race and especially by Hispanic ethnicity in the United States (Hellerstein and Neumark 2008); education explains very little of workplace segregation by race, but language skills explain about a third of Hispanic–white segregation. The racial segregation does not help explain lower black wages, as blacks actually work in higher-paying establishments. However, Hispanics tend to work in slightly lower-paying establishments.

2.2 *Age*

Older workers generally earn *more* than younger workers. However, it has long been suggested that older workers have greater difficulty finding new jobs. Older workers do have longer unemployment durations than middle-aged workers in US data, and prior to the 1967 Age Discrimination in Employment Act (ADEA), explicit age restrictions in hiring were common.[4] More recent evidence of such overt discrimination is documented in China and Mexico by Kuhn and Shen (2013) and Helleseter, Kuhn, and Shen (2014). During and after the Great Recession unemployment durations rose particularly sharply for older workers, as did ADEA claims filed with the US Equal Employment Opportunity Commission (EEOC) (Neumark and Button 2014).

2.3 *Evidence on Other Groups Covered in the Experimental Literature*

A number of studies document labor market differences for other groups that have been the focus of the experimental literature on discrimination. Researchers have documented earnings penalties for gay men, relative to straight men, of about 8 percent (with demographic, occupational, and other controls),[5] but earnings premia for lesbian women, relative to straight women (see Burn 2015). Many studies have documented earnings penalties associated with lower physical attractiveness (Hamermesh 2011). There is also evidence of wage

[2] See http://www.dol.gov/wb/stats/recentfacts.htm (viewed August 19, 2015).

[3] Occupational segregation may be more reflective of choice than discrimination, but firm-level segregation is harder to interpret as reflecting choice.

[4] For example, in five cities in states without anti-age-discrimination statutes, nearly 60 percent of employers imposed upper age limits (usually between ages forty-five and fifty-five) on new hires (US Department of Labor 1965).

[5] Since gays are identified for couples, based on living arrangement and relationship, the comparison is to married straight men.

penalties for obesity, even in longitudinal data with individual or family fixed effects (Baum and Ford 2004; Lundborg, Nystedt, and Rooth 2014). Wage penalties for obesity could also reflect factors similar to "lookism" (Caliendo and Gehrsitz 2015). Ameri et al. (2018) document that the employment rate for the disabled is less than half that of the nondisabled (33 percent versus 74 percent in 2012), their unemployment rate is about double (14.7 percent versus 7.2 percent in 2013), and they are paid less.

Finlay (2007) discusses descriptive evidence related to a criminal background, which points to lower wages and employment for those who have been incarcerated. However, the evidence is less clear conditional on individual fixed effects, consistent with negative selection into incarceration and less consistent with a causal effect. Pager (2003) also notes the strong and consistent negative relationships between incarceration and labor market outcomes, and that it is difficult to establish causal relationships, given the role of unobserved heterogeneity.

For at least some of these categories of workers, the differences in earnings and other outcomes can surely be less attributable to discrimination, for one of two reasons: the defining feature of the group may not be completely exogenous; or the defining feature of the group may in fact be associated with lower productivity. Of course, the discrimination literature more broadly considers the argument that differentials associated with sex, race, or ethnicity may reflect productivity differences, rather than discrimination. There may be a difference, however, for groups for which a productivity difference is in a sense *inherent* in the defining group characteristic, such as an employer concerned about liability for hiring a convicted felon, hiring a disabled worker in a job in which the disability lowers productivity, or who regards an employee's looks as contributing to their productivity.

3. *Explanations Based on Models of Discrimination*

In this section, I briefly discuss theoretical models of labor market discrimination.[6] The goal is to focus on what these models predict about possible findings from experimental studies—i.e., their observable implications—and what the models imply for how to interpret the experimental evidence.

3.1 *Taste Discrimination*

Becker (1957) considered three models that were distinct, but unified, in focusing on the market implications of discriminatory tastes of three different types of agents—employers, employees, or customers. The model of employer discrimination is the one through which the majority of empirical work on discrimination has been interpreted.

3.1.1 *Employer Discrimination*

In the employer discrimination model, employers dislike hiring a particular group, such as blacks.[7] When a black is hired, an employer considers the cost to be both the wage and the disutility from hiring the worker. Thus, we think of employers as utility maximizers, with utility function

$$U = U(\pi, B) = \pi - d \cdot B, \tag{1}$$

where π is profits, B is the number of black workers, and $d > 0$ is a constant, reflecting discriminatory tastes against blacks (a simple form of discriminatory tastes). Assume the production function is

$$y = f(W + B), \tag{2}$$

[6] Arrow (1998) provides an overview of the main theoretical models of labor market discrimination, including the potential limitations of these models.

[7] I use blacks and whites here to stand in for any two groups, the first of which might be disfavored. I change the reference to other groups when issues specific to those groups (e.g., men versus women or young versus old) arise.

a particularly simple form in which there are no inputs except labor, and W (whites) and B are perfect substitutes. The perfect substitutes assumption coupled with equal coefficients makes the meaning of discrimination clear, as whites and blacks are interchangeable in production.[8] We also assume that labor supply is perfectly inelastic, so wages are determined solely by the demand side. Normalizing the price of output to one, and letting w_i denote the wage of group i ($i = W, B$), the first-order conditions are

$$(3) \qquad MP_L = w_W, \quad MP_L = w_B + d.$$

Since d is positive, these imply $w_B < w_W$, and, in particular,

$$(4) \qquad w_W = w_B + d.$$

What does this imply about employment—and in particular hiring, which is the focus of experimental research on labor market discrimination? If d is the same for all employers, then when $w_W = w_B + d$ employers are indifferent between hiring white and black workers, so we expect most employers to hire both. In contrast, if $w_W \neq w_B + d$, employers will only want to hire one race or the other (e.g., only whites if $w_W < w_B + d$); as a consequence, wages will adjust in equilibrium to absorb both races, until $w_W = w_B + d$. In this case, there is no reason for employers to discriminate against blacks in hiring, as the wage differential just offsets the disutility from hiring. But if, as seems more likely, d varies across employers (d_j for employer j), an equilibrium wage differential should be established such that if $d_j > w_W - w_B$ the employer hires only whites, and if $d_j < w_W - w_B$, the employer hires only blacks. In this case, in an experimental study we would expect some employers to prefer hiring whites if there is discrimination.

However, things are a bit more complicated. In this formulation, employers with $d_j < w_W - w_B$ should prefer hiring blacks, so even if there are discriminatory employers, the implication for an experiment attempting to estimate a preference for black or white employees is unclear. If there is an equal wage constraint or at least a constraint on relative black and white wages (perhaps because of minimum wages), then it is more clear that we should, on net, find evidence of discrimination against blacks in hiring.[9] Alternatively, given that most wage equation evidence points to lower wages for blacks with similar observables to whites, we might presume that there must be a relative abundance of discriminatory employers, since even for the relatively small number of blacks to be absorbed by the labor market, the inframarginal employer would appear to be discriminatory.[10]

[8] This theoretical specification of discrimination is very close to one legal definition of discrimination in the United States. In particular, the Code of Federal Regulations (29, § 1604.2) defines as illegal discrimination "The refusal to hire an individual because of the preferences of coworkers, the employer, clients or customers …." However, note that statistical discrimination (discussed in the next section) is also illegal. The same regulations state: "The principle of nondiscrimination requires that individuals be considered on the basis of individual capacities and not on the basis of any characteristics generally attributed to the group." In an even more explicit reference to statistical discrimination as economists understand it, the same regulations note the following as illegal (i.e., not justifying a "bona fide occupational qualification" defense): "[t]he refusal to hire a woman because of her sex based on assumptions of the comparative employment characteristics of women in general. For example, the assumption that the turnover rate among women is higher than among men."

[9] Indeed, Neumark and Stock (2006) find robust evidence that sex discrimination/equal pay laws introduced in earlier decades (at the state and federal level) reduced the relative employment of black women and white women.

[10] A second complication is that this formulation of the employer discrimination model implies near-perfect employment segregation, with almost all firms hiring only one race or the other; but that is not generally observed. However, if we instead assume that employers care about the relative number of black employees, rather than the absolute number, then employers can alter the cost of hiring blacks by adjusting the ratio of blacks to whites, and

A different question is whether a finding of hiring discrimination in the kind of field experiment discussed below is actually informative about wage discrimination à la Becker. As Heckman (1998) emphasizes, the Becker model explores the market-level consequences of the presence of some employers with discriminatory tastes for equilibrium wage differences unrelated to productivity. But because wages are set at the margin, the presence of some discriminatory employers that an experiment on hiring discrimination might detect does not necessarily imply market-level discrimination in wages. In the formulation above, if there are enough nondiscriminatory employers relative to black workers that the former can employ all of the latter, then d at the marginal employer will equal zero, and black and white wages will be equal.[11] However, in search models with discrimination, the presence of any fraction of discriminatory employers that do not hire blacks can shift the offer wage distribution in the market, lowering equilibrium wages for blacks relative to whites (Black 1995).[12] Moreover, if these experiments are detecting statistical discrimination (discussed below), rather than taste discrimination, there is no reason the discrimination would not be reflected in market wages. Finally, individual-level discrimination based on characteristics of protected groups is illegal, so field experiments on hiring discrimination help us gauge whether policy has succeeded in rooting out this discrimination, or whether there are other forms of discrimination that policy makers might want to address.

Another question is whether employer discrimination can persist in the long run, or is undermined by competition. If discrimination cannot persist in competitive markets, and one believes markets are sufficiently competitive, then it is natural to be skeptical of claims of discrimination based on experiments—at least from field experiments in real markets. However, the claim that competition necessarily eliminates discrimination is often overstated. Even Becker (1957) clarified conditions under which employer discrimination could persist, and subsequent theoretical work further undermined the claim that competition would eliminate employer discrimination (Goldberg 1982).

3.1.2 *Employee and Customer Discrimination*

In the customer discrimination model, customers require a lower price when buying from blacks, because of the disutility cost of the transaction. Some of the experimental literature tests for customer discrimination by asking whether hiring discrimination is more severe for jobs with customer contact.

In the employee discrimination model, whites require a higher wage to work alongside blacks. This implies the following expressions for the profits of a firm in each of three scenarios: an all-white workforce, an all-black workforce, and an integrated workforce. These are, respectively:

$$\begin{aligned} (5)\quad \pi_W &= f(W) - w_W W, \\ \pi_B &= f(B) - w_B B, \\ \pi_{WB} &= f(W+B) - (w_W + d)W - w_B B. \end{aligned}$$

for any equilibrium wage ratio firms will be willing to hire some whites and some blacks (except for indivisibility problems).

[11] These conditions were elucidated in Becker (1957), and recently explored empirically by Charles and Guryan (2008).

[12] The wage-posting model of Lang, Manove, and Dickens (2005) has a similar result. In this model, even mild discriminatory preferences lead to an equilibrium with hiring discrimination and a wage difference in the market. Moreover, simply a belief among the minority group that firms discriminate—which can presumably be fueled by the experience of hiring discrimination at some employers—leads to the discriminatory equilibrium. Moreover, in this model no firm engages in wage discrimination à la Becker, paying a lower wage for a minority worker of equal productivity. Rather, wage discrimination emerges only from hiring discrimination.

Here, d is the wage premium that has to be paid to whites to work alongside blacks, and I use a simple specification in which they require d if they work with *any* blacks. This model predicts completely segregated workforces, with equal wages, and hence appears at first blush to be unable to explain the stylized facts, and has unclear implications for hiring discrimination experiments.

However, Arrow (1972) discusses extensions of the model that better fit the facts, and make it more likely that we would find lower hiring of blacks in a field experiment. The employee discrimination model implies, over some range, large quantity changes in response to small price changes. For example, starting from equal wages for whites and blacks, a slight decrease in the relative price of black labor should induce a firm with an all-white workforce to switch to an all-black workforce. This seems unlikely to occur, however, because of adjustment costs, and Arrow shows that once we account for adjustment costs, and presume that firms initially have predominantly-white or all-white workforces, it is easier to tell a story in which there is incomplete segregation of workforces, and blacks earn lower wages and have worse job opportunities. The same argument would apply to cases where we are considering discrimination against women, immigrants, etc., who may have arrived to or entered the labor market later.

3.2 *Statistical Discrimination*

In the simplest formulation of the statistical discrimination model (Aigner and Cain 1977), whites and blacks are assumed to be identical inputs on average, and there is individual-level variation in productivity that is unobserved by employers. Let q denote a worker's marginal productivity, with $w(q)$ the density of q for whites and $b(q)$ the density for blacks. Employers observe a noisy signal of q, $y = q + u$, and employers know that $E(u) = 0$, $\text{cov}(q,u) = 0$, $E(q) = \alpha$, and $\text{var}(u) = \sigma^2$. Employers form $E(q|y)$, and pay wages based on this. If q and u are jointly normally distributed, then

$$E(q|y) = \alpha \frac{\text{var}(u)}{\text{var}(q) + \text{var}(u)} + y \frac{\text{var}(q)}{\text{var}(q) + \text{var}(u)}. \tag{6}$$

The second ratio is the "reliability" of the signal, denoted γ. Average black and white wages are equal if $\alpha_W = \alpha_B$, so the statistical discrimination model does not generate the key stylized fact of unexplained wage differences groups. However, the model can be extended to obtain this result.

If $\gamma_B < \gamma_W$, it may be harder for employers to optimally match black workers to jobs, resulting in lower productivity and wages ex post. Alternatively, the noisier signal for blacks may deter blacks from investing in human capital, since their wages will be more tightly clustered around α, lowering $E(q)$ and therefore wages (Lundberg and Startz 1983). Another possibility is that initially erroneous beliefs about differences in α can become self-fulfilling (Coate and Loury 1993). In the case of men and women, for example, employers may initially assume that expected tenure is lower for women than for men, and therefore choose to invest less in women's human capital; this can give women less incentive to stay with the firm, leading to employers' expectations being fulfilled, and lower wages for women.

The statistical discrimination model is not tied to prejudice. Rather, especially for versions of the model that generate wage differences, it is best viewed as a model of stereotyping based on assumed group averages. Thus, the model can be directly related to many lab experiments that attempt to measure stereotyping. The model may not imply differences in hiring if wages are free to adjust to expected productivity differences.

If there are constraints on wage differences between groups, statistical discrimination should generate hiring differences that work against groups with lower assumed average productivity. In addition, if matching workers' abilities to the job is important, a noisier signal can lead to more adverse outcomes.

The distinction between statistical discrimination and taste discrimination is important for policy. In response to taste discrimination, the most natural policy response is to raise the cost to the employer of engaging in discriminatory behavior to, effectively, try to restore equal prices for equally productive labor from different groups. In contrast, policy interventions that increase information or the reliability of information about workers may reduce statistical discrimination.[13,14] This is one reason experimental research tries to distinguish between these two broad models of discrimination.

3.3 *Long-Term Incentive (Lazear) Contracts*

Lazear's (1979) model of long-term incentive contracts (LTICs) provides an explanation for behavior that looks like age discrimination with regard to both terminations and hiring of older workers. In this model, incentive-compatible contracts entail paying young, low-tenured workers less than their marginal product, and older, high-tenured workers more than their marginal product. The model offers an explanation of mandatory retirement, which is predicted to occur when the discounted stream of wage payments catches up to the discounted stream of marginal productivity; mandatory retirement was viewed as discrimination by the ADEA, which abolished it over time. LTICs may also deter hiring of older workers by imposing fixed costs that can be amortized only over a shorter period for older workers (Hutchens 1986); barriers to paying newly hired older workers much lower wages than current older workers to create the needed incentives can lead to the same result. Finally, LTICs can encourage employers to renege on implicit contracts, terminating higher-tenure workers when their wages exceed marginal product. Does the differential treatment of older workers predicted by this model represent discrimination? Clearly statistical discrimination with regard to how long workers will stay with the firm plays a role. Moreover, differential treatment based on the Lazear model has been interpreted as discrimination from a legal perspective (Issacharoff and Harris 1997).

3.4 *Implicit Discrimination*

The final explanation of discrimination that is also tested in the experimental literature is "implicit discrimination." The basic idea is that discriminatory decisions—making decisions based on race, for example, when race is not actually predictive of differences in productivity—are unconscious rather than conscious. Devine (1989) provides an overview of social psychology research on the difference between conscious efforts that may actively combat stereotypes that can lead to discriminatory behavior, and unconscious actions or decisions in which these stereotypes operate even among decision makers who are not prejudiced or who deliberately try to avoid stereotypes or prejudice.[15]

[13] As an example, Holzer and Neumark (2000) suggest that affirmative action may have precisely this effect, encouraging more investment in scrutiny of minority workers both before and after hiring, and reduced reliance on the most readily available signals.

[14] One possible exception is that a less noisy signal, when there is a wage floor that limits paying wages as low as some workers' expected productivity, can lower hiring.

[15] Greenwald and Banaji (1995) provide a broader overview of the development of the idea of unconscious or implicit cognition—including but not limited to stereotyping race and other groups; the paper also discusses a large body of empirical work.

Bertrand, Chugh, and Mullainathan (2005) describe the implicit association test (IAT), which claims to measure this kind of unconscious discrimination, and argue that scores on the IAT predict other behaviors that may translate into discriminatory behavior.[16] Some experimental research tries to assess whether implicit discrimination generates discriminatory hiring differences.

If implicit discrimination is important in hiring and other decisions, then standard policy responses that operate on conscious behavior, such as reinforcing the illegality (and potential financial consequences) of discrimination, may not be very effective. In contrast, unconscious biases may be more amenable to policy interventions (including at the employer level) that reduce the role of unconscious decision making—such as formalizing evaluation procedures, or things as simple as allowing more time to evaluate job candidates.

4. *Nonexperimental Approaches to Testing Hypotheses about Discrimination*

Economists have used a number of nonexperimental approaches to test for and estimate the extent of labor market discrimination, and to try to distinguish among the different models of discrimination. Here, I highlight key issues that arise in the nonexperimental literature that help illuminate what experimental methods can provide, and also touch on where nonexperimental approaches offer advantages relative to experimental work.

4.1 *Regression Decompositions*

The traditional approach to studying labor market discrimination focuses on the role of discrimination in generating wage gaps between groups. The approach entails estimating wage regressions, trying to control for productivity-related differences between two groups, and interpreting the remaining wage gap between the groups as an estimate of discrimination, most commonly using separate regressions to allow wage differences between groups to vary with other characteristics (e.g., Oaxaca 1973; Neumark 1988). The challenge that the regression decomposition approach addresses, of course, is that the groups in question often have differences in numerous characteristics that could be related to productivity—such as differences in labor market attachment between men and women, or in school quality between blacks and whites—so that unadjusted differences in wages need not reflect discrimination.[17]

Although regression adjustments can, in principle, account for these differences, unexplained wage gaps can always be attributed to unmeasured productivity-related characteristics for which the regression fails to account, rather than discrimination. This points to the principal advantage of experimental approaches, which instead create artificial pools of labor market participants among whom there are supposed to be no average differences by group, so that observed outcomes are more convincingly attributed to discrimination. A second potential problem with the regression decomposition approach is that the regression controls can reflect responses to discrimination, such as lower experience and tenure of women because of discrimination

[16] The reader can take these tests at https://implicit.harvard.edu/implicit/education.html (viewed September 1, 2015). For a skeptical discussion of the validity of the IAT as a predictor of discriminatory behavior (or as a better predictor than explicit measures)—in many cases in laboratory studies—see Oswald et al. (2013). Steffens (2004) presents evidence on ability to fake the IAT.

[17] Darity and Mason (1998) provide an interesting overview of relatively recent literature using this approach, especially with regard to race and the incorporation of test scores into such regressions.

on the job (e.g., Gronau 1988)—sometimes termed "feedback" effects—which can lead to understating discrimination by conditioning on its effects. Some recent natural experiment studies (in the context of sports) have led to interesting new evidence of feedback effects.

A potential advantage of the regression decomposition approach is that it speaks directly to wage differences between groups. In contrast, experimental studies focus mainly on hiring, which (as discussed above) may or may not have implications for wages depending on the underlying model, the size of the minority group, and the share and distribution of discriminatory tastes. Experimental studies have, for the most part, not provided direct evidence on the effects of discrimination on wages.

4.2 *Production Function Tests*

More recently, labor economists have estimated productivity differences directly, and compared these differences to wage differences. Hellerstein and Neumark (1999) developed a method of using matched employer–employee data to do this. In an application using US data, Hellerstein, Neumark, and Troske (1999) jointly estimate plant-level production functions and wage equations, and find a sex gap in wages that significantly exceeds the sex gap in marginal productivity, consistent with sex discrimination. This structural approach to estimating wage discrimination has now been used in a number of countries where matched data have become available (see Bartolucci 2014).

The production function approach has the advantage, relative to the regression decomposition approach, of direct measurement of worker productivity (although since productivity is measured at the plant level, group differences are identified from across-plant variation in productivity and the share of workers in different groups). In that sense, the production function approach quite clearly provides more compelling evidence on wage discrimination. On the other hand, it is difficult to fully rule out a potential role for sorting of workers across plants (although enhancements to the approach by Bartolucci (2013) try to do this). The experimental methods discussed below generally avoid this sorting problem, since identical applicants are evaluated by potential employers (or proxies for them in the lab). Again, though, the experimental approach is generally focused on hiring, not wages.

4.3 *Tests of Statistical Discrimination*

A few nonexperimental studies test for statistical discrimination, and sometimes try to distinguish between statistical and taste discrimination. These studies illustrate some of the same difficulties I emphasize below with regard to experimental studies that try to distinguish between taste and statistical discrimination. In particular, distinguishing between these models in nonexperimental studies requires strong informational assumptions regarding what is observed by firms but potentially correlated with unobserved productivity, and how the information available to firms changes over time.

The tight relationship between tests for statistical discrimination and what employers know about workers, and when they know it, was first made explicit in Altonji and Pierret's (2001) research on the evolution of black and white wages as workers gain experience, to test for statistical discrimination based on race. It is perhaps demonstrated most clearly in Foster and Rosenzweig (1993), who study data from the Philippines that include demographic characteristics and wages paid for time-rate and piece-rate work. Given that piece-rate work should capture differences in productivity, we might test for taste discrimination against women by regressing time rates (w) on a piece-rate measure of

productivity (μ) and a dummy variable for women (F):

$$w = \alpha + \beta\mu + \gamma F + \varepsilon. \tag{7}$$

Indeed, Foster and Rosenzweig's data show that w is lower for women ($\hat{\gamma} < 0$), conditional on μ. However, if employers do not know individual-level productivity when they set time rates, but just know that μ is lower on average for women than for men, then the estimated coefficient on μ is biased toward zero, and the average difference can load onto the female dummy variable, making statistical discrimination look like taste discrimination. If we could measure expected productivity, $E(\mu)$, and there is no taste discrimination but only statistical discrimination, the coefficient on $E(\mu)$ should equal one, and the coefficient on F should equal zero. But if we still find a negative coefficient on F, conditional on $E(\mu)$, this would point to taste discrimination.

Since there is no proxy for $E(\mu)$, Foster and Rosenzweig (1993) instrument for μ, assuming that μ is the sum of $E(\mu)$ and a random forecast error. Valid instruments are variables that employers use in forecasting μ, but that do not affect wages directly—precluding, for example, taste discrimination based on these variables. As instruments, the authors use height, age, and education. The resulting instrumental variables estimates do not show any evidence of lower wages paid to women conditional on $E(\mu)$, consistent with all of the initial difference in the ordinary least squares (OLS) estimates being attributable to statistical discrimination. Note that this approach hinges critically on assumptions about what employers know and when they know it—in particular, that employers do not know productivity μ when they set time rates, but that they observe other characteristics that predict productivity.[18]

[18] One way to mitigate the strong reliance on such assumptions is to use exogenous variation in how much employers know about workers. As one example, Finlay (2007) studies statistical discrimination in hiring from groups more likely to have a criminal background, estimating how the advent of criminal background checks may have enabled employers to better distinguish among individual blacks, reducing statistical discrimination against blacks without a criminal record. A similar approach is used—in reverse—by Agan and Starr (2018), who study the impact of recent initiatives to ban employers from using very early elicitation of criminal background information in recruitment.

There are some direct parallels to experimental studies, which try to test for statistical versus taste discrimination by adding information that employers might use to predict differences between worker productivity and seeing whether this changes outcomes. However, the challenge regarding who knows what about unobserved productivity, and when, carries over to this experimental research, although as I discuss below, the problem has not been recognized as explicitly. On the other hand, I suggest that it may be possible to augment experimental studies with survey or other evidence that can help answer these questions about what information employers have, and when they have it and hence perhaps lead to more definitive tests.

5. *Laboratory Experiments*

Lab experiments have been used to test for discriminatory decision making, to try to understand the nature of discrimination, and to study behaviors that can lead to discriminatory outcomes. There is a broader research literature on laboratory experiments that explores alternative, nondiscriminatory explanations of group differences in outcomes, like recent work on gender and competitiveness (e.g., Gneezy, Niederle, and Rustichini 2003). There is also an extensive psychology literature measuring stereotypes about different groups (e.g., Kite et al. 2005). And there are lab experiments on discrimination in settings other than labor

markets (like the public goods experiment in Andreoni and Petrie 2008).

I restrict attention to research that focuses on discrimination, especially research using laboratory settings meant to mimic potential labor market discrimination. Thus, I exclude research on gender and competitiveness, the role of race, etc., in other outcomes (like public goods experiments), and the general literature on gender and bargaining except where it tries to address labor market outcomes directly. And I cover research on stereotypes only to the extent that this research ties the stereotypes participants hold to actions that could plausibly be tied to labor market outcomes. My goal is not to cover every study, but to illustrate the kinds of studies that have been done, to discuss the issues that arise, and to identify important questions. And where survey articles are available, I reference and discuss them, rather than going back to the individual studies, and offer my perspectives on the larger literature they cover.[19] Most of the studies I discuss are summarized in table 1.

There are standard objections regarding laboratory versus field experiments, mostly having to do with external validity and with capturing realistic stakes for participants.[20] The latter issue may arise in a particularly complex way in research on discrimination because the market-level effects of competition may be the driving force that imposes costs for engaging in discriminatory behavior, implying that it may be hard to mimic the costs of discrimination in the laboratory. In fact, though, a good deal of the existing laboratory experiment research does even less to incentivize behavior, imposing no actual costs on participants to make discriminatory choices in selection of employees in hypothetical scenarios. Such considerations might increase the evidence of discrimination in the laboratory. On the other hand, it is possible that respondents in the lab are more prone to make socially desirable choices that could mask discrimination (Norton, Vandello, and Darley 2004; Norton et al. 2006).[21] The absence of cost-related incentives may be especially problematic with respect to statistical discrimination, which can be profit maximizing for real-world employers and hence detected in field experiments, but may be masked in lab experiments where social desirability is influential and there are no costs to foregoing statistical discrimination.

5.1 *Simulated Personnel Decisions* (*Vignette Studies*)

Vignette studies examine discrimination in the laboratory by trying to simulate personnel decisions made by employers or managers. Researchers present participants with hypothetical scenarios regarding selection of job candidates for hiring, or of employees for training, promotion, etc., and often subsequently elicit attitudes toward these workers or other supplemental information that might help explain participants' decisions. As in the field experiments discussed below, protected-group status (age, sex, etc.) is manipulated in these studies for otherwise identical job candidates, which is what makes vignette studies experiments.

It is easy to criticize vignette studies for dealing in hypotheticals. However, for personnel decisions regarding existing workers (rather than new hires), it is difficult to conceive of a corresponding field experiment, such as selecting otherwise identical

[19] The same applies to section 6, on field experiments.

[20] On the second issue, see, e.g., Niederle and Vesterlund (2007) and Flory, Leibbrandt, and List (2015).

[21] Mujcic and Frijters (2013) present interesting evidence of this in a different context. In their experiment (in Australia), black, white, Asian, and Indian testers tried to get free rides on buses (claiming their bus pass was faulty, which it was). Blacks and Indians were granted free rides at a far lower rate. But in a survey of bus drivers presented with such a scenario, the black and Indian differences were actually reversed slightly.

TABLE 1
SELECTED RESULTS FROM LABORATORY EXPERIMENTS ON LABOR MARKET DISCRIMINATION

Study	Method	Summary/Results
Sex discrimination		
Rosen and Jerdee (1974)	Vignettes administered to bank managers in training program regarding personnel decisions.	*Discrimination against women.* Higher selection of men than women for promotion and professional development. Higher likelihood of approving termination in response to request from male supervisor when performance is the issue.
Olian, Schwab, and Haberfeld (1988)	Meta-analysis of vignette studies.	*Discrimination against women in selection.* Similar negative effect sizes for student and professional subjects.
Correll, Benard, and Paik (2007)	Vignettes administered to undergraduates for rating and selection, same-sex married pairs that differ by parenthood status.	*Discrimination against mothers among female, married women.* Large motherhood penalty, and smaller fatherhood "premium," in rating, recommended salary, and recommendation for hiring. Hiring outcome mediated but not eliminated by controls for ratings. No sex discrimination among nonparents.
Ethnicity discrimination		
Blommaert, Coenders, and van Tubergen (2014)	Vignettes administered to students in university and higher vocational education institutions.	*Discrimination against minorities.* Weakly lower evaluations for Turkish or Moroccan ethnicity versus Dutch, and stronger evidence of less selection for interviews. Discrimination in selection for interviews weaker for respondents reporting positive interethnic contacts.
Age discrimination		
Rosen and Jerdee (1977)	Vignettes administered to subscribers of *Harvard Business Review*, ages 32/61.	*Discrimination against older workers.* Respondents more likely to recommend reassignment from older worker in response to complaints, and less likely to recommend selection for development or promotion. Stereotypes indicate older workers perceived as less likely to change behavior, less likely to want to keep up with technology, and less likely to succeed in position requiring creativity and innovation.
Rosen and Jerdee (1976)	Vignettes administered to undergraduate business students, ages 32/61.	*Discrimination against older workers.* Respondents more likely to recommend reassignment from older worker in response to complaints, and less likely to recommend selection for development or promotion. Stereotypes indicate older workers perceived as less likely to change behavior, less likely to want to keep up with technology, less likely to succeed in position requiring creativity and innovation, and less desirable for a financial position involving quick judgments and high risk.
Weiss and Maurer (2004)	Vignettes administered to undergraduate engineering students, replication of Rosen and Jerdee (1976), ages 32/61.	*Little evidence of age discrimination.* Few differences found between younger and older workers, with the only exception that older workers were deemed less likely to change behavior in response to complaints, prompting a higher likelihood of recommending the work be reassigned.
Büsch, Dahl, and Dittrich (2009)	Vignettes administered to students and personnel managers in Germany and Norway.	*Evidence of age discrimination.* Declining selection with age in Germany, where participation by older workers is lower than in Norway. Conditional on a suggested wage range, which may reflect productivity, evidence of age discrimination in both countries.

(*Continued*)

TABLE 1
SELECTED RESULTS FROM LABORATORY EXPERIMENTS ON LABOR MARKET DISCRIMINATION (*Continued*)

Study	Method	Summary/Results
Looks and obesity discrimination		
Hosoda, Stone-Romero, and Coats (2003)	Meta-analysis of vignette studies.	*Discrimination against less attractive.* Large positive advantage for the attractive. Does not vary much with the sex of the applicant, the sex type of the job, the provision of job-relevant information, or whether subjects were students or professionals. Effect declines over time.
Rudolph et al. (2009)	Meta-analysis of vignette studies.	*Discrimination against obese.* Adverse effects of obesity for performance evaluation, selection for promotion, and hiring (for which the effects are largest).
Statistical versus taste discrimination		
Heilman (1984)	Vignettes administered to MBA students for selection for managerial position. Sex manipulated, as is provision of information on college coursework.	Lower selection of women (and ratings of potential and likelihood of success) when less job-relevant information is provided. Difference largely eliminated when job-relevant information provided (coursework in economics and business) that may undermine sex stereotypes.
Tosi and Einbender (1985)	Meta-analysis of vignette studies of selection by sex.	Across studies, sex differences in outcomes are lower when subjects are given more information about applicants.
Davison and Burke (2000)	Meta-analysis of vignette studies of selection by men and women into "opposite-sex" jobs.	For both men and women, lower ratings for opposite-sex jobs are diminished when more information is provided. Effect may have more to do with reduced salience of sex, rather than statistical discrimination.
Baert and De Pauw (2014)	Vignettes administered to students regarding ethnic discrimination, with additional survey questions.	Subjects were more likely to indicate that customers or coworkers would enjoy interacting with natives more than minorities, although they indicated that as employers, there is no difference in selection for hiring. Subjects were likely to rate the minority applicants as having the required productivity for the job, but to rate the minority group overall as less qualified. Interpreted as evidence of customer and coworker discrimination, but not employer statistical discrimination; validity of interpretation can be questioned.
Mobius and Rosenblat (2006)	Lab experiment in which workers solve mazes (ability unrelated to beauty) and employers interact with workers in different ways (some visual and some not) to estimate productivity and pay workers.	Employers overestimate ability of better-looking workers, forming incorrect stereotypes. No role for taste discrimination.
Bargaining		
Dittrich, Knabe, and Leipold (2014)	Male and female subjects alternate roles as employer or employee, with same- and mixed-gender pairings.	*Evidence of wage discrimination in bargaining framed as employer–employee negotiations.* Males acting as employees negotiate higher wages than females, mainly because of higher initial wage offers of males as employees, and higher initial wage offers of male employers to male employees. The differences emerge because of differences in initial offers rather than subsequent bargaining, suggesting that sex differences in bargaining skills are not critical.
Implicit discrimination		
Reuben, Sapienza, and Zingales (2014)	Lab experiment in which male or female employers hire male or female workers for mathematical task.	*Evidence of negative stereotypes against women in mathematical tasks, and limited updating based on information.* Male and female participants are much more likely to hire males than females with no information about the ability of the potential hires. Bias is greater, and degree of updating based on performance information is less, for those with stronger implicit biases.

employees for training or promotion decisions. The only exception would be, perhaps, an experiment run internally at a company that was sufficiently large to be able to present fictitious workers to other managers.[22] The closest laboratory experiments have gotten to more real-world decisions is to administer vignette studies to managers likely to be actively involved in making similar decisions—for example, administering the experiment to managers participating in a training program. To a large extent, vignette studies simply test for discrimination in decision making, without attempting to determine the nature of discrimination. However, some of the studies present evidence that could be viewed as testing statistical versus taste discrimination.

There are scores of vignette studies in the literature. Here, I touch on some that are cited frequently and that appear to be most closely related to hypotheses that arise in the economics of discrimination.[23]

5.1.1 *Sex*

An early vignette study (called, at the time, an "in-basket exercise") was administered to male bank managers participating in a management training program (Rosen and Jerdee 1974). The managers were given hypothetical scenarios that required selecting bank employees for promotion or development, and the solution of a supervisory problem based on the report of a supervisor whose sex was manipulated. The study found significantly lower willingness to promote females, whether into complex or simple jobs, and significantly lower selection of females for development (attending a professional training conference). In supervisory problems, a lengthy report from a supervisor described a problem and requested that the employee be terminated or at least transferred; the choices of participants ranged from terminating the employee, to arranging a transfer, or "softer" solutions. Termination was more likely to be recommended when the supervisor was male and the workplace issue concerned performance (with the evidence less clear when the issue was personality). All of this evidence is consistent with sex discrimination in decisions related to key labor market outcomes. The authors interpret the evidence as confirming the importance of sex stereotypes in sex discrimination, but the basis for this conclusion is not clear.

Researchers have considered the external validity of these vignette studies, albeit in the limited sense of asking whether outcomes using professionals instead of students as subjects yield different results. In an early meta-analysis, Olian, Schwab, and Haberfeld (1988) found a similar (negative) effect size of female for studies using the two kinds of subjects, although for some reason a higher variance of effect sizes for professionals. This may partially mitigate concerns that outcomes for student subjects are not reflective of real-world decision making, although what professionals do in the lab versus the real world remains an open question.

The studies reviewed in this section are dated, and vignette studies of sex discrimination appear to be less common in more recent literature. However, more recent experimental research has explored questions that intersect with gender. For example, Correll, Benard, and Paik (2007) conduct a laboratory experiment and a field experiment (discussed below) of the motherhood penalty, motivated by evidence suggesting that much

[22] Baert, De Pauw, and Deschacht (2016) try to study promotions in the context of a correspondence study field experiment, but this entails hiring at new employers into jobs at a higher level than the present one, which is different from internal promotions.

[23] As counterexamples, Singer and Sewell (1989) discuss many studies regarding selection decisions based on age. However, many of these studies pertain to cognitive issues (the impact of other information calling attention to age) or sociological issues (the role of differences in the status of jobs considered).

of the unexplained contemporaneous sex gap in wages is attributable to motherhood. In the lab experiment, undergraduate subjects evaluate same-sex pairs of married applicants who differ in whether they are parents. For the female pairs, ratings of competence and commitment are significantly lower for mothers, as are recommendations for hiring (47 percent versus 84 percent, significant at the 5 percent level) and recommended salaries. For the male pairs, the recommended hiring rate is a bit higher for fathers (73 versus 62 percent, significant at the 10 percent level), as is recommended salary. The differences in evaluations of competence and commitment account for about 40 percent of the hiring differential within female pairs. Finally, there was no evidence of lower ratings or selection for hiring between non-parent men and women (based on between-pair comparisons).

5.1.2 *Ethnicity*

More recently, vignette studies have been applied to ethnic discrimination, coupled with additional evidence on potential determinants, or at least correlates of discrimination. Blommaert, Coenders, and van Tubergen (2014) present student participants in the Netherlands with resumes signaling Moroccan, Turkish, or Dutch ethnicity via name and parents' country of birth (all applicants were born in the Netherlands), as well as sex, education, and experience. Participants rated applicants (from zero to ten) on suitability for the job, and selected three they would invite for an interview. Minority applicants received lower suitability ratings, although the effect was very small (0.05 of the zero-to-ten ranking, yet still statistically significant). For selection for interview, the odds ratio for selecting native Dutch applicants was 1.14, marginally significantly different from one.

Participants were surveyed on personal and background characteristics. Indeed, the authors argue that the ability to interview or survey experiment participants is an advantage of laboratory versus field experiments of discrimination—although, below, I will discuss field experiments that try to collect information from recruiters. Blommaert, Coenders, and van Tubergen (2014) find less discriminatory behavior in selection for interviews by those who report more positive interethnic contacts (higher quality, not higher frequency), as well as those with higher education. The first result might provide some support for a taste-based understanding of ethnic discrimination in this experiment, although there is no way to know whether positive interethnic contacts reduce discriminatory tastes, or are simply correlated with weaker discriminatory tastes.

5.1.3 *Age*

Rosen and Jerdee (1977) conducted a vignette study of age discrimination similar to their earlier sex discrimination study, based on a survey administered to *Harvard Business Review* subscribers. They provided hypothetical scenarios and asked subscribers how they would respond to managerial complaints about workers, decisions about career development, and promotions. Age of the worker in question was manipulated (32 or 61). In response to complaints about a worker, respondents indicated greater difficulty in changing the behavior of older workers, and were more likely to recommend that the work be reassigned from an older worker than from a younger worker, and conversely more likely to recommend addressing the issue via an encouraging talk for younger workers. In response to a request for a career development opportunity regarding production technology, older workers were viewed as less likely to want to keep up with technology, and the decision was more likely to be approved for a younger worker. And in response to promotion to a position requiring

creativity and innovation, older workers were viewed as less likely to succeed, and were less likely to be recommended for promotion. Similar results are reported in Rosen and Jerdee (1976), using as subjects, instead, undergraduate business students.

However, twenty-eight years later, a replication of the latter study by Weiss and Maurer (2004) found relatively little evidence of negative stereotypes or adverse hypothetical decisions for older workers. This could be attributable to more older workers in the workforce and changes in attitudes and behaviors over three decades. It would be preferable to know about changed attitudes and behaviors among subjects like subscribers to the *Harvard Business Review* (as in Rosen and Jerdee 1977). Although these papers provide suggestive evidence that age-related stereotypes are likely to affect managerial decisions, the analyses do not tie individual responses regarding stereotypes to the recommendations of those respondents.[24]

Büsch, Dahl, and Dittrich (2009) carry out a similar type of study for white-collar jobs that are not physically demanding, using as participants both students asked to assume they were assistants of a personnel manager, and actual personnel managers contacted by mailing questionnaires (with low response rates). The authors did the study in both Norway and Germany, hypothesizing that age discrimination would be worse in Germany, in part because of lower participation of older workers in the labor market. Artificial applicants were ranked on twelve capabilities. I focus on the personnel manager responses, which are likely more informative about actual labor market behavior. The Norwegian managers did not rate older workers lower on any capabilities (point estimates were similar for young, middle-aged, and older applicants), whereas in Germany older workers were ranked lower (than middle-aged or younger workers) on capability to learn and flexibility. This was reflected in the hypothetical hiring decisions as well, as the selection rate declined significantly with age in Germany.[25]

5.1.4 *Looks and Obesity*

Hosoda, Stone-Romero, and Coats (2003) provide a meta-analysis of laboratory studies of discrimination based on looks or "attractiveness" and selection for promotion, performance evaluation, hiring, and other outcomes. Attractiveness is typically manipulated by including a picture, although there are differences, such as having subjects rate mock videotaped job interviews (Riggio and Throckmorton 1988). Hosoda et al. conclude that "Physical attractiveness is always an asset for individuals" (p. 451). The effect size is large (although a bit hard to interpret), and the effect does not appear to vary with the sex of the applicant, the sex type of the job, the provision of job-relevant information, or whether subjects were students or professionals. However, the effect size declines over time (across studies), and is largest for choice of a business partner and smallest for performance evaluation.

Rudolph et al. (2009) provide a similar type of meta-analysis of studies of discrimination against obese job applicants. The authors note that obesity can be associated with unattractiveness, moral or emotional impairment, unhappiness, higher absenteeism, and lower acceptance by coworkers, emphasizing the issue of whether studies

[24] Krings, Sczesny, and Kluge (2011) study selection by age and stereotypes regarding age and "warmth" and competence (older workers are rated higher on the former, and lower on the latter). The authors found lower selection of older workers, and that this bias was mediated by age-related stereotypes regarding competence.

[25] An interesting twist was that the respondents were also asked to indicate an appropriate wage range for the candidate. Conditioning on these wage ranges increased the evidence of discrimination in selection against older workers.

capture differences in discrimination or productivity. The manipulation of weight is sometimes done using computer morphing programs (Polinko and Popovich 2001) or theatrical prostheses (Pingitore et al. 1994) to make the same "applicant" appear as both average weight and obese. The evidence from this meta-analysis points to adverse effects for evaluation and selection for hiring.

5.2 *Statistical versus Taste Discrimination*

Some laboratory studies try to estimate whether providing additional information about applicants reduces differences in selection between groups. Such evidence might be interpreted as testing between statistical and taste discrimination, although this interpretation is made more explicit in field experiments taking this approach, and hence gets more attention when I discuss field experiments.

Heilman (1984) documents inconsistent findings in past vignette studies regarding whether providing additional information about individual applicants reduces adverse outcomes (in selection) for women. Heilman suggests that the discrepancies can arise because sex is a salient characteristic for evaluators, so that additional job-relevant information will reduce the role of sex, but information not relevant to the job can worsen outcomes for women by increasing the salience of sex. This is tested in an experiment on selection of a recent college graduate for a managerial position. The "job-relevant" information treatment is success in courses in economics and business. The "low-relevance" treatment is biology and political science. (In the third treatment, neither type of information is given.) The results are consistent with her conjectures. In particular, only the provision of job-relevant information eliminates the lower selection of women, potentially consistent with statistical discrimination.

The role of job-relevant information in reducing group differences is considered in three meta-analyses. Tosi and Einbender (1985) report that, *across* studies, sex differences in selection and other outcomes are smaller when subjects are given more information about applicants. Davison and Burke (2000) focus on ratings of both sexes for "opposite-sex" jobs and similarly report that lower ratings for opposite-sex jobs are diminished when more information is provided. The evidence that information affects ratings of both males and females in this way suggests that the findings may reflect the reduced salience of sex, rather than statistical discrimination per se (which we might expect to disadvantage women). In contrast, Hosoda, Stone-Romero, and Coats (2003) also report across-study results, finding that providing more job-relevant results about applicants does not influence group differences in outcomes based on attractiveness.

Three points can be made about these studies. First, we might expect less statistical discrimination based on looks, since it is not clear what unobserved productivity differences employers might associate with attractiveness. Second, it is hard to know how to interpret evidence that more information fails to reduce group discrimination (as in Hosoda, Stone-Romero and Coats 2003, but more generally), since we do not know whether the information provided pertains to qualifications about which employers make assumptions based on group averages when discriminating statistically; this is a general problem with testing taste versus statistical discrimination, to which I return below. Finally, it is clearly more desirable to study variation within rather than across studies, to isolate the effects of providing information.

A related problem of thinking about the role of information in interpreting studies like these arises in Baert and De Pauw (2014), who conduct a laboratory experiment on discrimination in Belgium against applicants with Turkish (versus Flemish) names, using students as subjects. They find

that respondents were likely to rate minority applicants as having the required productivity for the job (indeed, more likely than nonminorities), although respondents rate the minority group population overall as less qualified. The authors interpret this as inconsistent with statistical discrimination, because the overall lower group average for the minority group is not reflected in evaluation of the candidates. However, the experimental design presents native and minority candidates with identical resumes, so subjects may have enough information that they do not impute the group-level averages to the applicants they assess. That is, thinking back to the Aigner and Cain (1977) framework, we have no way of knowing whether subjects perceived the signals of applicant quality as sufficiently noisy that they would put a lot of weight on group averages of the applicant populations. Absent this information, it is hard to see how the evidence can be informative about statistical discrimination.[26]

In my view, however, the external validity problem is particularly severe with regard to these kinds of tests for statistical discrimination. To infer something about real-world statistical discrimination from whether providing additional information on hypothetical applicants affects group differences in selection requires assuming that the laboratory conditions regarding information actually mimic real-world hiring. In the case of Heilman's analysis, for example, it seems unlikely that employers *would not* know the field of study to begin with, so the experimental treatment of no information about coursework and how the outcome changes when this information is provided may not provide evidence on how employers' actual decisions are changed by the provision of information that might reduce sex stereotypes.

In contrast, Mobius and Rosenblat (2006) conduct an experimental study of the "beauty premium" that more successfully distinguishes between taste discrimination and statistical discrimination. The experiment involves solving mazes, for which there is no difference in ability based on looks. "Workers" have to estimate how many mazes they can solve, and "employers" have to estimate how many mazes workers can solve (both types are incentivized to estimate correctly). Each worker is matched with five different employers, and the interaction is manipulated to include no visual or oral (by telephone) information, one type information or the other, both types, or face-to-face interaction. Wages paid to workers are based on some of the employer estimates, and come out of employers' endowments. Payments are arranged so that the employer first evaluates or interacts with the worker, and is then told which estimates will contribute to a worker's earnings. This provides evidence on taste discrimination, since an employer might sacrifice earnings to pay a higher wage to an attractive worker.

The authors decompose the beauty premium into a number of sources. They find that one contributor is overconfidence regarding better-looking workers, and another is overestimation of ability of better-looking workers by employers. The latter is a form of statistical discrimination, albeit based on incorrect stereotypes, because better-looking people are in fact no more productive in their setting. There is no evidence of taste discrimination, although the authors caution that this kind of result may not carry over to real-world labor markets, where the utility gain from interacting with a good-looking person over an extended period may matter much more than in the lab—a question, again, of realistic stakes and external validity.

[26] Moreover, since the study's results provide no evidence of (hypothetical) employer discrimination in the first place (there is no difference in selection for hiring), it is hard to interpret this study as testing alternative explanations of such discrimination.

5.3 *Tests for Implicit Discrimination*

There has been little or no work trying to test whether implicit discrimination can account for group differences in labor market related outcomes in laboratory experiments.[27] Rosen and Jerdee (1977) found that although respondents exhibited age discrimination in decisions and held age stereotypes, a large share favored policy reforms that would help older workers, such as complete vesting of pensions and the elimination of mandatory retirement. They suggest that these two types of responses/behavior are discrepant, and could imply that respondents in the role of managers do not consciously discriminate, but respond to unconscious stereotypes—which sounds like an early appeal to implicit bias. However, this inference may not be supported. First, only some respondents (60–80 percent) supported policies to help older workers, and the others may have driven the age differences in personnel decisions and stereotypes. Second, the responses concern policy reforms that are not closely related to the personnel decisions respondents were asked to make.

More recently, Reuben, Sapienza, and Zingales (2014) conduct a laboratory experiment in which "employers" hire "workers" to do a mathematical task. They find that participants in the role of employers are much more likely to hire males than females when they have no information about the ability of the potential hires. This pattern is reduced, but not eliminated, when information on prior performance on the mathematical task is provided. Finally, the researchers also administer an IAT based on associating men or women with words more related to math/science versus liberal arts, and find more bias and less updating based on performance data for those who, according to the IAT, have stronger implicit biases.

[27] Dovidio et al. (2002) survey experimental psychology research on implicit discrimination.

5.4 *Discrimination in Bargaining Experiments*

There is experimental research on sex differences in bargaining outcomes (e.g., Eckel 2008; Solnick 2001), which of course could influence pay. Some of the work on bargaining also pertains to taste versus statistical discrimination (Fershtman and Gneezy 2001). Here, though, I restrict the discussion to a bargaining experiment that more explicitly tries to capture features of labor markets.[28]

Dittrich, Knabe, and Leipold (2014) study multistage, alternative-offer bargaining, which they believe is more likely to generalize to labor market outcomes than simple bargaining experiments, and they explore differences depending on whether males and females are playing the role of employees or employers. Their experiment is framed as a firm bargaining with a worker. The firm needs a task done, for which it earns a fixed amount, and proposes a wage to the worker, the remainder of which is the firm's earnings. The worker has an outside option of zero, and the employer has a positive outside option that is less than its total earnings if it reaches an agreement. An initial offer is made, alternatively by the employer or the employee, and then there are rounds of counteroffers. After the third round, if there is no agreement reached there

[28] There are also field experiments on bargaining, but perhaps not surprisingly, given the needs of this research, these experiments focus on consumer transactions. Examples include sex differences in bargaining over taxi fares in Peru (Castillo et al. 2013), race, age, and sex differences in bargaining over sports cards (List 2004), and sex and race differences in bargaining over car prices (Ayres and Siegelman 1995). These studies do not pertain to the labor market, so I do not cover them. But it is interesting to note that they often point to statistical discrimination as the source of differences in outcomes (e.g., assumptions about the valuation of a taxi ride in the Castillo et al. study). Taste-based discrimination might be more important in labor markets because employment entails a long-term relationship.

is an exogenously imposed probability of the negotiations ending, in which case each gets their outside option. Pairings can be either mixed or same gender, and those bargaining sit across from each other but communicate only via a fixed script.

There are two key findings. First, males acting as employees negotiate higher wages than females (but as employers, there are not any differences in outcomes). Second, male employers make higher initial wage offers to male employees than female employees, whereas female employers do not behave in this way. The authors interpret this finding as wage discrimination. Third, the differences emerge because of differences in initial offers (of male employees generally, and male employers to male employees), rather than subsequent bargaining, suggesting that sex differences in bargaining skills are not critical.[29]

6. *Natural Experiments*

A handful of studies on discrimination in labor markets can be considered natural experiments. In some, variation in the race of who evaluates "workers" is effectively random, giving us information on how different workers are evaluated depending on the characteristics of the evaluator. And in one, the evaluation mechanism exogenously changed from evaluators knowing the sex of applicants to sex-blind evaluations, which is informative about discrimination prior to the advent of sex-blind selection.[30]

By way of analogy, suppose that, as in a traditional nonexperimental study of wage discrimination, we had data on workers' wages, race, sex, and human capital controls. However, suppose we also had repeated observations on the race of the supervisor who was setting wages, and this varied for reasons exogenous with respect to pay. Alternatively, suppose we had evidence that in some cases pay was set without the supervisor knowing the worker's sex, again for exogenous reasons. These are not realistic scenarios, yet we would likely regard them as natural or at least "quasi" experiments. The handful of studies I discuss in this section, which are summarized in table 2, have the same flavor.

6.1 *Sports*

Kahn (2000) describes how detailed data on performance, work histories, "supervisors," and pay in the world of sports can be used to study discrimination. Some recent papers have pushed the envelope in this area of research to consider natural experiments. These studies perhaps do not get at the most fundamental questions, such as who gets hired and how much they are paid. But they use clever research designs that provide clean evidence on how workers are evaluated, and how this affects performance.

Parsons et al. (2011) study the evaluation of Major League Baseball pitchers by umpires, looking at pitches near the edges of the strike zone, where there is more uncertainty as to whether a pitch is a ball or a strike. They assemble a large data set in which pitchers and umpires appear multiple times, and are matched randomly—based on both the data and the institutional assignment of umpires

[29] In a related paper that asks whether bargaining differences explain sex differences in earnings, Exley, Niederle, and Vesterlund (2016) develop a bargaining experiment intended to ask whether women fare worse by avoiding negotiating—i.e., by not "leaning in"—couching the experiment in terms of workers and firms splitting revenue. The evidence suggests that men and women fare similarly when forced to negotiate, but when given a choice, women avoid negotiation more. However, women's unwillingness to negotiate does not harm them often. Instead, women appear to positively select into negotiating when it will pay off, whereas men do not.

[30] Indeed, as discussed in the next section, there are a couple of field experiments where a similar change was made.

TABLE 2
RESULTS FROM NATURAL EXPERIMENTS ON LABOR MARKET DISCRIMINATION

Question/issue	Study	Method	Summary/Results
Sports			
Do umpires favor pitchers of the same race/ethnicity in making ball and strike calls in Major League Baseball?	Parsons et al. (2011)	Estimate whether in same-race/ethnicity pairings of umpires and pitchers, strikes are more likely to be called, and measures difference when there is more explicit or implicit monitoring.	*Discrimination when players evaluated by umpires of different race/ethnicity.* Strikes are more likely to be called when the umpire and pitcher are of the same race/ethnicity. Bias affects pitchers' behavior so that performance of other-race pitchers is worse. The bias in umpiring diminishes considerably under more intensive monitoring.
Are more fouls called on players of the opposite race by referees in the National Basketball Association?	Price and Wolfers (2010)	Estimate relationship between racial composition of referee crew and race of player in calling fouls.	*Discrimination when players evaluated by referees of opposite race.* Fouls are more likely to be called on players whose race deviates more from that of the umpiring crew. The bias mainly affects white players, either via worse treatment by black referees or favorable treatment by white referees. Bias appears to affect players' behavior adversely.
Orchestras			
Did the switch to blind auditions for major orchestras increase the selection of females?	Goldin and Rouse (2000)	Regression models for selection of females, using variation in the adoption by orchestras of blind auditions, where the musician plays behind a screen and other steps are taken to ensure that the musician's identity is not known when selection decisions are made.	*Discrimination against women in selection*. In general, the selection of females increased as a result of blind auditions, and this change in audition procedures accounted for about one quarter of the increase in the percentage of females in the period studied.

to games. They can therefore estimate models conditional on pitcher and umpire fixed effects and obtain causal estimates of how the race or ethnicity of the umpire and the player affects strike calls. Their evidence indicates that strikes are more likely to be called when the umpire and pitcher are of the same race/ethnicity. Moreover, umpire evaluations affect pitcher performance. A different-race pitcher is less likely to pitch to the edges of the strike zone, where they will experience this disadvantage, and this strategic response affects the pitcher's success adversely, because a pitch to the middle of the strike zone is easier to hit.

The study also exploits variation in monitoring of umpires—either explicitly through the use of technology or implicitly based on crowd size and the strategic importance of the pitches. They find that the bias in umpiring disappears under more intensive monitoring, and appears only (and more strongly) in the low-monitoring environments. Parsons et al. interpret the evidence as indicating that when the "cost" of discrimination is higher, umpires engage in less of it. This might be suggestive of implicit discrimination.

A study of calling of fouls by referees in the National Basketball Association (Price and Wolfers 2010) provides evidence of a similar nature. The authors know the racial composition of the three-person refereeing crew and the race of players (they focus on blacks and nonblacks). Using similar analyses as Parsons et al. (2011) (although with no evidence of explicit monitoring, and perhaps less clear

evidence of implicit monitoring), they find that fouls are more likely to be called on players whose race deviates more from that of the refereeing crew, and that this bias appears to affect players and teams adversely. Interestingly, this bias mainly affects white players, either via worse treatment by black referees (in contrast to what we think generally might happen in the labor market) or favorable treatment by white referees.[31]

The evidence in these two studies, indicating that discriminatory biases in evaluation can affect performance, means that performance measures in discrimination studies cannot necessarily be treated as exogenous. That is, if discrimination leads to worse performance (as seems likely in these contexts, but not necessarily in all contexts),[32] regressions of pay or related measures on performance and group membership will understate discrimination by overcontrolling.[33] This is, in a sense, another manifestation of the kinds of feedback effects that have been discussed in nonexperimental research on discrimination. Theory also suggests how statistical discrimination (which is probably not what the sports studies are detecting) can lead to behavioral responses that can mask discrimination—such as "self-fulfilling" negative stereotypes stemming from employers' initial erroneous beliefs about group differences (Coate and Loury 1993). However, note that in the audit and correspondence field experiments discussed in section 7, there is no reason to expect feedback responses that could diminish measured discrimination, since artificial applicants are made identical.

6.2 *Orchestras*

Goldin and Rouse (2000) study how the switch to blind auditions for major orchestras affected the selection of female musicians. The variation in this study arises from the adoption by orchestras, over time, of blind auditions where the musician plays behind a screen and other steps are taken to ensure that the musician's identity is not known when selection decisions are made. The authors find that the selection of females increased because of blind auditions, suggesting that there was discrimination against women prior to the adoption of blind auditions.[34] The contribution of this study to understanding discrimination in the labor market is more substantial than the sports studies, because the audition process relates directly to hiring. Indeed, Goldin and Rouse cast their study as complementary to the kinds of audit or correspondence studies of discrimination discussed in the next section of the paper.

The study includes a cautionary tale about how either natural or experimental variation

[31] Price and Wolfers (2010) suggest that their results could reflect implicit discrimination. For a related application in sports that claims to test for implicit discrimination, see Gallo, Grund, and Reade (2013). This research can be viewed as trying to exploit variation in the conditions under which implicit discrimination is expected to occur, which can potentially provide evidence on implicit discrimination without relying on the IAT.

[32] In contrast, think of the Charlotte Whitton quote: "Whatever women do they must do twice as well as men to be thought half as good," (famously followed by "Luckily, this is not difficult").

[33] There is related evidence in a recent study by Glover, Pallais, and Pariente (2017), who find that when minority cashiers are paired with more biased managers (as measured by the IAT), they perform relatively worse (e.g., they scan items more slowly, and take more time between customers) than nonminority cashiers paired with those managers. Earlier evidence of these kinds of feedback effects on job interview performance is presented in Word, Zanna, and Cooper (1974).

[34] The use of blind auditions is chosen by the orchestra, and hence the variation is not really "natural." The authors are careful to rule out orchestras that hire more women ex ante adopting blind auditions (or adopting them earlier)—analogous to testing the parallel-trends assumption in a quasi-experimental panel data analysis. The study also illustrates the importance of controlling for worker heterogeneity (via a fixed effects analysis that evaluates outcomes for workers under different treatments), because the adoption of a more neutral screening procedure that reduced sex discrimination increased the likelihood that less-qualified female musicians auditioned.

can induce behavior that eliminates *reverse* discrimination. In particular, Goldin and Rouse (2000) discuss a semifinal stage when affirmative action may have been used to increase the pool of women, which can no longer occur when auditions are blind. Below, I discuss a couple of field experiments where adopting sex- or race-blind procedures seems to have reduced hiring opportunities for women or minorities, presumably by reducing affirmative-action-like favoritism. In the Goldin and Rouse study, however, this presumably unintended consequence at one stage of the process was less important than the larger overall impact of blind auditions in increasing the selection of women.

7. *Field Experiments on Hiring Discrimination*

The most extensive body of experimental research on labor market discrimination is audit or correspondence (AC) studies of discrimination in hiring. There are lengthy discussions of the basic methods elsewhere (Fix and Struyk 1993), and a few survey papers or meta-analyses that provide broad overviews of the findings for some groups. I focus on key issues in the design of field experiments and analysis of data from them to help the reader understand how these studies have developed and to try to identify unresolved or potentially promising research questions. I also provide a more detailed overview of results for groups that have not been the focus of the bulk of the existing studies and, hence, of the prior surveys.

7.1 *Basic Methods*

AC studies of the labor market are field experiments used primarily to address the question of discrimination in hiring. In audit studies, fake job candidates ("testers") of different races, ethnicities, etc., who are sometimes actors, are sent to interview for jobs (or in some early studies, apply by telephone). The candidates have similar resumes and are often trained to act, speak, and dress similarly. Correspondence studies, in contrast, use fictitious job applicants who exist on paper only (or now, electronically), and differ systematically only on group membership. The response captured in correspondence studies is a "callback" for an interview or a closely related positive response. In contrast, the final outcome in audit studies is actual job offers. The earliest AC studies of discrimination were Daniel (1968)—an audit study—and Jowell and Prescott-Clarke (1970)—a correspondence study.[35] AC studies have also been used to analyze discrimination in housing markets and consumer markets. My focus is on labor markets and the unique issues that arise in these markets.[36]

It is easy to see the appeal of experimental AC studies. Nonexperimental regression-based approaches to testing for and measuring discrimination use data on the groups in question in a population, introducing regression controls to try to remove the influence of group differences in the population that can affect outcomes. AC studies, in contrast, create an artificial pool of labor market participants among whom there are supposed to be no average differences by group. This is clearly a potentially powerful strategy because if we have, e.g., a sample of blacks and whites who are identical *on average*, then in a regression of the form

$$Y = \alpha + \beta B + \varepsilon, \tag{8}$$

[35] Schwartz and Skolnick (1962) did an early version of a correspondence study of the effects of criminal background. But this study did not focus on a characteristic usually associated with discrimination (e.g., race).

[36] The application to housing markets originated with the US Department of Housing and Urban Development's Housing Market Practices Study in 1977. For an overview of some of the existing work, see http://www.urban.org/features/exposing-housing-discrimination (viewed November 22, 2015). For recent research on consumer markets, see Gneezy, List, and Price (2012), Doleac and Stein (2013), and Zussman (2013).

where Y is the outcome and B is a dummy variable for blacks, ε is uncorrelated with B, so that the OLS estimate $\hat{\beta}$ (or simply the mean difference in Y) provides an estimate of the effect of discrimination on Y.[37]

For some of the discussion that follows, a more formal framework is useful.[38] Suppose that productivity depends on two individual characteristics (standing in for a larger set of relevant characteristics), $X' = (X^I, X^{II})$, so that productivity is $P(X')$. Characteristic X^I is observed by firms, and X^{II} is unobserved. It is simplest, for now, to think of Y as continuous, such as the wage offered, although in AC studies we should think of it as latent productivity leading to a decision to hire/call back or not. Define discrimination as

$$(9) \quad Y(P(X'), B = 1) \neq Y(P(X'), B = 0).$$

Assume that $P(\cdot, \cdot)$ is additive, so

$$(10) \qquad P(X') = \beta_I X^I + X^{II},^{39}$$

and

$$(11) \qquad Y(P(X'), B) = P + \gamma B.$$

Discrimination against blacks implies that $\gamma < 0$, so that blacks are paid less than equally productive whites.

In AC studies, researchers create resumes that standardize the productivity of applicants at some level. Denote expected productivity for blacks and whites, based on what the firm observes, as P_B^* and P_W^*. Note that Y is observed for each tester, so each test—the outcome of applications to a firm by one black and one white tester—yields an observation

$$(12) \quad Y(P_B^*, B = 1) - Y(P_W^*, B = 0) = P_B^* + \gamma - P_W^*.$$

Given that the AC study design sets $P_B^* = P_W^*$, we should be able to estimate γ easily from these data, by simply running a regression of Y on the dummy variable B and a constant.[40] However, there are potential complications.

Nearly all studies used matched pairs of testers (or even triplets or quadruplets), usually justified as a means of controlling for heterogeneity across employers (indeed one can include employer fixed effects in the regression models estimated).[41] In a recent paper, Phillips (forthcoming) suggests that some existing studies appear to understate discrimination because of positive spillovers that arise from matched-pair comparisons in which the quality of the artificial applicants submitted influences employers' evaluations of other applications submitted (although the bias can go in either direction).[42] Such

[37] Of course, most of the regression studies focus on wages, whereas AC studies focus on hiring. In section 3.1.1, I discussed why hiring discrimination detected by AC studies might have implications for wage discrimination.

[38] This is a simplified and abridged version of Neumark (2012).

[39] Because X^{II} is unobservable, its coefficient is normalized to one.

[40] Moreno et al. (2012) present an interesting analysis using observational data from a government job intermediation service in Peru. They are able to obtain detailed data on applicants for jobs and characteristics of jobs. They refer to this as a "pseudo-audit" study, because it uses actual rather than fictitious applicants.

[41] A recent exception is Weichselbaumer (2015a), who studies discrimination in Germany against women with Turkish names wearing headscarves (in Germany, photos are included with job applications). She sends only one application to each employer, for two reasons: first, she argues that correspondence studies have gotten enough publicity in Germany that employers may look for closely matching applications; and second, job applications in Germany require extensive amounts of information, including detailed school reports, making it hard to match within pairs. The basic statistical analysis is the same, but there can be differences in the clustering of observations and/or the fixed effects that can be included.

[42] See also related evidence in Weichselbaumer (2015b). Although based on small samples, she finds weaker evidence of discrimination in a paired-application

spillovers are unlikely if the number of artificial applications is small relative to the number of applications the employer receives, but typically researchers do not know the latter number. For researchers concerned with this issue, the problem can be avoided by forgoing matched-pair designs—for example, randomizing race for each applicant independently within pairs—so that the "treatment" of being assigned a black name is not correlated with the applicant pool composition (as affected by the researcher).

With matched-pair designs, some researchers discard pairs of testers in which neither gets a callback or interview and measure "net discrimination" as the difference between the number of pairs where the white is favored over the black versus the opposite, divided by the number of tests with at least one positive response (e.g., Riach and Rich 2002). However, recent work often translates the results of AC studies into differences in how many jobs a member of a group has to apply to in order to get a callback or interview (e.g., Bertrand and Mullainathan 2004; Rich 2014), for which the number of pairs in which neither gets a response is clearly relevant. Moreover, the full data seem critical to gauging discrimination. For example, suppose that there are one hundred tests in a study, and in ninety-seven neither tester gets a positive response, in two the white tester does, and in one the black tester does. Discarding the pairs with no offers implies net discrimination of 33.3 percent, whereas a more reasonable interpretation would be that it is only slightly harder for a black to get a positive response than a white. Standard practice has become to report estimates from linear probability or probit models (marginal effects), retaining all the observations; this approach does not require matched pairs.

design, which she conjectures may be because employers may detect the test and avoid or reduce discrimination.

7.2 *Data and Design Issues*

There are a number of decisions researchers have to consider in designing an AC study.[43] The first is whether to do an audit or a correspondence study. A major advantage of audit studies is that actual job offers are observed, which are more convincingly tied to labor market outcomes, although, perhaps surprisingly, audit studies have not proceeded to the stage of wage offers.[44]

A second issue in this choice is that callback rates could be influenced by factors that make them less reliable indicators of ultimate hiring. Suppose that a correspondence study uses highly skilled white and black candidates. Employers may know that the representation of highly skilled blacks in the population is lower. Hence, if they want to hire some blacks, they may make a disproportionately high number of callbacks to blacks, expecting more competition for them from other employers, which could boost callback rates for blacks and obscure evidence of discrimination.[45] However, some audit studies on ethnic discrimination by the International Labor Organization (ILO), discussed in Riach and Rich (2002), separate discrimination at the selection for interview and job offer stages, and find that most (around 90 percent) of the discrimination occurs at the selection for interview stage, suggesting that callbacks for interviews constitute the key part of the hiring process to

[43] Readers interested in designing and conducting a correspondence study might find the detailed discussion and "roadmap" in Neumark, Burn, and Button (forthcoming) useful. Here, I touch on the main issues with less of a "how-to" flavor.

[44] Another potential advantage noted by Pager (2007) is that in audit studies, testers can be debriefed regarding their treatment during the interview, providing some qualitative evidence on what happened, although this can inject a subjective element.

[45] Of course, this can also happen for job offers in audit studies, since not all offers are accepted.

study.[46] Moreover, nearly all correspondence studies find evidence of discrimination against minorities, so at worst, callback studies could be understating racial and ethnic discrimination, rather than obscuring it. In contrast, the evidence of sex discrimination from correspondence studies is weaker and more ambiguous. However, it is unlikely that higher callbacks per female applicant relative to intended hiring are responsible for this evidence, because skill distributions are more similar between men and women. Nonetheless, there are still open questions of what we learn from differences in callback rates in correspondence studies, relative to job offers in audit studies.

On the other hand, correspondence studies have many advantages. One is the much lower cost per application, which permits researchers to obtain larger samples, with which they can explore a richer set of questions. As examples, the large-scale Urban Institute audit studies of discrimination against blacks and Hispanics (Cross et al. 1990; Turner, Fix, and Struyk 1991) had observations numbering only in the hundreds. In contrast, Bertrand and Mullainathan's (2004) correspondence study of black-sounding names had over 5,000 observations, and Neumark, Burn, and Button's (forthcoming) correspondence study of age discrimination had over 40,000 observations. Both correspondence studies (and especially the latter) exploit these large samples to design variations in resumes to test alternative hypotheses about discrimination.

Correspondence studies also let researchers avoid experimenter effects that can lead to bias in audit studies. For example, Heckman and Siegelman (1993) point out, in their evaluation of the Urban Institute audit studies, that part of the training of testers was "a general discussion of the pervasive problem of discrimination in the United States" (p. 216), raising the possibility that testers took actions in their job interviews that led to the "expected" result. Similarly, Pager (2003) reports that black testers posing as having criminal records may have had negative psychological reactions that affected their performance in interviews. Correspondence studies avoid this problem, effectively creating blind testers.[47]

More generally, in audit studies the influence of small differences between testers from different groups can be amplified by the nature of the study design. For example, Heckman and Siegelman (1993) note that Hispanic testers in one of the Urban Institute test sites had facial hair. Any productivity difference associated with facial hair might seem trivial compared to the productivity-related characteristics employers care about. But when the research design matches on many productivity-related characteristics, the importance of these trivial features can be magnified. This problem is avoided in correspondence studies because resume characteristics are randomized.[48] For the preceding

[46] Similarly, in my audit study of sex discrimination in restaurant hiring (Neumark 1996), we were able to compare callbacks resulting from candidates dropping off resumes in the morning, when managers were not there and there was little personal interaction, to results for job offers based on the actual interviews. The evidence pointed to discrimination at both the invitation for interview and job-offer stages.

[47] Indeed, Jowell and Prescott-Clarke (1970) carried out the first correspondence study, rather than using what they called "actor applicants," because of concerns about motivation: "It has been suggested that there could well be conscious or unconscious motivation on the part of coloured actors to *prove* discrimination, and that this would bias the results" (p. 399).

[48] One way to test for the influence of uncontrolled tester characteristics in audit studies is to use multiple pairs of testers and explore the robustness of the findings across different pairs of testers, perhaps using tests of the homogeneity of results for the different pairs prior to pooling the data (Heckman and Siegelman 1993). For example, in my audit study of sex discrimination in restaurant hiring (Neumark 1996), the two female testers were quite different looking, and one had more success in getting offers at high-price (and high-pay) restaurants, so we verified the robustness of the overall results to using only pairs with the more successful female tester.

reasons, the correspondence-study method has come to dominate field experiments on labor market discrimination, and hence much of the ensuing discussion of methods focuses on correspondence studies.

The second key design issue—which undergirds the quality of the research project—is the construction of applicants and their resumes. One decision concerns which kinds of jobs to study. Typically, these studies focus on lower-skill, entry-level jobs, for three reasons: first, the ways in which job applications are made and responses collected (e.g., by email) are more common for low-skill jobs;[49] second, websites now list numerous low-skill jobs to which researchers can apply; and third, for higher-skill jobs there is likely a greater chance that candidates would be known to prospective employers or that employers could easily learn something that reveals the applicant as fictitious.[50] Aside from missing higher-skill jobs, correspondence studies likely miss jobs found through informal contacts and referrals, which are important job search methods (Ioannides and Datcher Loury 2004).

A key goal in constructing applicant resumes is to make them realistic for the jobs being sought. Recent studies have used websites on which resumes are posted as the basis for the resumes used in the study (e.g., Bertrand and Mullainathan 2004; Neumark, Burn, and Button forthcoming).[51] The devil is in the details here; Neumark, Burn, and Button (forthcoming) provide a recent discussion of many of the potential issues that arise in constructing resumes and suggest how to use data, as much as possible, to resolve them.

In correspondence studies, group membership has to be signaled on the resume. This is probably easiest with regard to sex, since most first names are not gender-neutral, or with respect to age, since many resumes list year of high school graduation (Neumark, Burn, and Button forthcoming). Race is not listed on resumes, so Bertrand and Mullainathan (2004) use typically black and white names based on birth certificate data (e.g., from their title, "Lakisha and Jamal" versus "Emily and Greg"), as have more recent race studies. This introduces the possibility that employers associate productivity differences with black-sounding names rather than race per se. Bertrand and Mullainathan try to rule this out by controlling for neighborhood quality.

Other characteristics are even less natural to signal on resumes. For example, studies of discrimination based on sexual orientation (e.g., Drydakis 2009) have included a line on some resumes indicating volunteer work on behalf of the gay community or a gay organization, versus including similar work for a different kind of organization (e.g., environmental) on other resumes, to signal sexual orientation while making the resumes neutral with respect to activism or volunteering. In their study of disability, Ameri et al. (2018) disclose one of two disabilities in a cover letter.

There are potentially interesting questions as to how different ways of signaling group membership may affect outcomes, perhaps most importantly by unintentionally providing information on productivity differences. The issue of black-sounding names was already discussed. As another example, Figinski's (2017) correspondence study of military reservists wrestles with whether selection into the military, or past activations, are associated with productivity-related factors. Similarly, Ameri et al. (2018) try to choose disabilities that are

[49] On the other hand, Zschirnt and Ruedin (2016) note that correspondence studies might not cover the very lowest-skill jobs where written or online applications are not used.

[50] There are exceptions to a focus on low-skill jobs. In a recent study of discrimination against the disabled, Ameri et al. (2018) used fictitious applicants who were either novice or experienced accountants, and had callback rates in the 5–8 percent range, only a shade lower than, for example, the 6–10 percent range for lower-skilled sales and administrative jobs in Bertrand and Mullainathan (2004).

[51] Sometimes actual resumes posted on job-search websites are used (e.g., Siddique 2011).

unlikely to affect productivity (suggesting that one of the disabilities they study—Asperger's syndrome—may even be associated with higher productivity in accounting).

7.3 *Differences in Unobservables*

A correspondence study can preclude systematic differences between groups in observables. But there can still be differences in employers' assumptions about unobservable differences between groups. I discuss, in turn, differences in levels and differences in variances, which raise different issues.

7.3.1 *Differences in Means of Unobservables*

As noted above, the group difference in outcomes we estimate in a correspondence study is

$$(13)\quad Y(P_B^*, B=1) - Y(P_W^*, B=0) = P_B^* + \gamma - P_W^*.$$

Using B and W subscripts in the obvious way on X^I and X^{II}, $P_B^* = E(\beta_I X_B^I + X_B^{II} | X_B^I, B=1)$, and similarly for P_W^*. Assuming randomization, and with $X_B^I = X_W^I = X^I$, equation (13) reduces to $\gamma + E(X_B^{II}|X^I, B=1) - E(X_W^{II}|X^I, B=0)$, implying that we only identify γ if $E(X_B^{II}|X^I, \ B=1) = E(X_W^{II}|X^I, B=0)$. But employers may have different expectations about the mean of X^{II} for blacks and whites, conditional on what they observe, which a labor economist would label statistical discrimination.

Thus, absent any method of distinguishing between γ and $E(X_B^{II}|X^I, B=1) - E(X_W^{II}|X^I, B=0)$, correspondence studies identify the combined effects of taste discrimination and statistical discrimination.[52] However, if the study provides much less information about applicants than employers usually have, then the statistical discrimination that is identified as part of the sum $\gamma + E(X_B^{II}|X^I, B=1) - E(X_W^{II}|X^I, B=0)$ might have little to do with real-world statistical discrimination. On the other hand, when a correspondence study includes a rich set of applicant characteristics, it becomes less likely that statistical discrimination plays much of a role in group differences in outcomes. These issues come to the fore in studies that try to distinguish between statistical and taste discrimination.

7.3.2 *Differences in Variances of Unobservables*

A second problem is that employers may assume different *variances* of the unobservables across groups. With data on hiring—the focus of AC studies—it is most natural to think of applicants having to exceed some productivity threshold with sufficiently high probability, and in such threshold models, different variances of the unobservable can matter. I refer to this problem, described in Heckman and Siegelman (1993) and Heckman (1998), as the "Heckman critique" of AC studies. The possibility of a difference in the variance of unobservables is in fact a central feature of early models of statistical discrimination (Aigner and Cain 1977).

To isolate the problem, consider the best-case scenario where $E(X_B^{II}|X^I, B=1) = E(X_W^{II}|X^I, B=0)$, so there is no statistical discrimination regarding levels. But the standard deviations of the unobservables, denoted σ_B^{II} and σ_W^{II}, need not be equal.[53] Assume the applicant is called back (hired) if there is a sufficiently high probability that their productivity exceeds a given threshold. In this case, $\sigma_B^{II} \neq \sigma_W^{II}$ generates bias in the

[52] Indeed, if implicit discrimination provides a different reason for undervaluing the productivity of a group of workers in AC studies, the empirical implication for the framework developed here would likely be the same as the implication of taste discrimination.

[53] I assume homoskedasticity within groups and hence suppress conditioning on X_B^I and X_W^I.

estimate of discrimination; worse, perhaps, we cannot necessarily even sign the bias.

To see this intuitively, suppose that the research design standardizes X^I at a low level, denoted X^{I*}. Employers care about how likely it is that the sum $\beta_I X^I + X^{II}$ exceeds some threshold. Given the low value X^{I*}, this is more likely for a group with a high variance of X^{II}. Thus, even in the case of no discrimination ($\gamma = 0$), the employer will favor the high-variance group. Conversely, if standardization is at a high level of X^{I*}, the employer will favor the low-variance group. Because researchers do not have information on the population of real applicants to the jobs studied, there is no definitive way to know whether X^{I*} is high or low relative to the actual distribution, and hence no way to sign the bias.

I suggested a solution to this problem that can separately identify the relative variances of the unobservables and the discrimination coefficient γ (Neumark 2012).[54] The intuition behind the solution stems from the fact that a higher variance for one group (say, whites), ceteris paribus, implies a smaller effect of observed characteristics on the probability that a white applicant meets the standard for hiring. Thus, information from a correspondence study on how variation in observable qualifications is related to employment outcomes can be informative about the relative variance of the unobservables, and this, in turn, can identify the effect of discrimination. Based on this idea, the identification problem identified by the Heckman critique is solved by invoking an identifying assumption—specifically, that the effect of applicant characteristics that affect perceived productivity, and hence callbacks, have equal effects across groups—along with the testable requirement that some applicant characteristics affect the callback probability (since if all the effects are zero we cannot learn about $\sigma_B^{II}/\sigma_W^{II}$ from these coefficient estimates).

In a probit specification, for example, we know that we can only identify the coefficients of the latent variable model for productivity relative to the standard deviation of the unobservable. In this case, we effectively have two probit models, one for blacks and one for whites. If we normalize σ_W^{II} to one, then for a characteristic (Z) that affects the callback rate, we identify its coefficient (δ_W) relative to σ_W^{II}, or δ_W/σ_W^{II}. However, if we assume that $\delta_W = \delta_B$, then we do not need to impose the normalization that $\sigma_B^{II} = 1$, but instead can identify $\sigma_B^{II}/\sigma_W^{II}$ from the ratio of the coefficients on Z in the probit for whites versus blacks, which in turn allows us to identify γ. The estimation can be done using a heteroskedastic probit model, which also permits an approximate decomposition into the effect of race via γ and its effect via the difference in variances of the unobservable. Finally, when there are *multiple* productivity-related characteristics that shift the callback probability Z_k ($k = 1, \ldots, K$), there is an over-identification test because the ratio of coefficients on each Z, for whites relative to blacks, should equal $\sigma_B^{II}/\sigma_W^{II}$.

It is actually rare that correspondence studies include variables that shift the callback probability, because these studies typically create one "type" of applicant for which there is only random variation in characteristics that are not intended to affect outcomes. However, a handful of studies have this information—including Bertrand and Mullainathan's (2004), whose data I used to illustrate this approach. Moreover, studies can be designed to include this kind of quality variation.

This discussion raises the issue of what we are trying to measure in AC studies. To this point, I have focused on γ, which I think

[54] To reiterate, in this section I assume $E(X_B^{II} | X^I, B = 1) = E(X_W^{II} | X^I, B = 0)$, to simplify. Without this assumption, references to γ in the remainder of this section should be read as references to $\gamma + E(X_B^{II} | X^I, B = 1) = E(X_W^{II} | X^I, B = 0)$—i.e., the sum of taste and statistical discrimination.

of as the structural effect of race capturing the potential discounting by employers of black workers' productivity à la Becker as in equation (11) (and possibly statistical discrimination about the mean of X^{II}). But we just saw that employers could treat blacks and whites differently in hiring because of different variances of the unobservable. Should differential treatment because of unequal variances be interpreted as discrimination, in which case we might not want to eliminate it?

The reason the coefficient γ is the object of interest is that the taste discrimination (and possibly "first-moment" statistical discrimination) that AC studies capture in γ generalizes from the correspondence study to the real economy. In contrast, in the Heckman and Siegelman framework, the "second-moment" statistical discrimination described above is an artifact of how a correspondence study is done—i.e., the standardization of applicants to particular, and similar, values of the observables, relative to the actual distribution of observables among real applicants.[55] If, instead, a study used applicants that replicated the actual distribution of applicants to the employers in the study, there would be no bias—in the setting described here—from the different variances of the unobservable; that is, the differential treatment that is detected is an artifact of the study design. That said, the correction for bias is wedded to a specific model of hiring as well as a distributional assumption. There is no general reason to expect that this method would eliminate bias in the estimate of discrimination that might arise in a model in which higher-order moments affect hiring decisions directly, or with non-normal distributions.

[55] That is not to say there cannot be discrimination based on second moments with, for example, risk averse firms (see, e.g., Dickinson and Oaxaca 2009, for lab experiment evidence).

7.4 *Testing the Nature of Discrimination*

A number of recent AC studies try to distinguish whether taste or statistical discrimination underlies their evidence of hiring discrimination. Note that in both field and laboratory experiments, the statistical discrimination that implicitly underlies these tests has to do with lack of employer information and the assumptions employers may make about worker characteristics they do not see in the information provided. The issue of whether group differences in outcomes are arising from group differences in characteristics that arise endogenously from discrimination (as in Coate and Loury 1993) is not at issue, because the studies eliminate these group differences. Thus, the approach to testing for statistical discrimination is generally to add information to resumes, and if doing so diminishes the difference in callback rates between groups, the evidence is interpreted as indicating that employers must have been statistically discriminating—i.e., assuming group differences with respect to the types of information provided in the "high-information" treatment.

There are two potential problems with this approach. First, we do not know, ex ante, on what unobserved characteristics employers might be statistically discriminating. If we add information that is not the basis for employers' statistical discrimination, then a null finding of no change in callback rates is uninformative. Second, a reduction in the difference between callback rates from adding information to the resumes does not necessarily imply statistical discrimination, because the experiment—and in particular the "low-information" treatment—may not mimic actual employer hiring. To take an extreme case, suppose we compare results using resumes with no information except the group identifier to resumes with other information (like education and job histories) that employers would typically have, and

we find that the callback rate difference between the groups diminishes. This does not imply that, in the real world, members of the disadvantaged group suffer from statistical discrimination, because employers would actually have the information on education and job histories. A valid test requires that we know what information is typically *not* provided in the job application process, on the basis of which employers statistically discriminate, and examine the effect of adding that information.

To see that there is no "general" result that adding information to an AC study is informative about whether statistical discrimination generates group differences in hiring or callback rates, using the notation from above, suppose that

$$E(P|B=1) < E(P|B=0). \tag{14}$$

Then in an audit study where employers know only race (B), in the model for hiring

$$Y = \alpha + \beta B + \varepsilon, \tag{15}$$

the estimate β reflects both taste discrimination (γ) plus the expected productivity differential. If the resumes instead also provide information on X, and

$$E(P|B=1,X) = E(P|B=0,X), \tag{16}$$

then the estimate of β reflects only taste discrimination, and hence the effect of adding X to the resumes might be interpreted as providing evidence of statistical discrimination. But if real-world employers know X, then this evidence is only an artifact of using an unrealistic baseline "low-information" treatment that excludes X.

Conversely, suppose we know that employers have information on X but not on Z, and the low-information treatment includes X while the high-information treatment also includes Z. If

$$E(P|B=1,X) < E(P|B=0,X), \tag{17}$$

but

$$E(P|B=1,X,Z) = E(P|B=0,X,Z),$$

then adding Z reduces group differences from statistical discrimination. But if

$$E(P|B,X,Z) = E(P|B,X), \tag{18}$$

that is, Z is uninformative about productivity, then adding Z has no impact on estimated group differences.

Thus, the validity of this kind of test for statistical discrimination in AC studies hinges, in part, on how well the job applications in the low-information treatment capture the information employers actually have, and—less knowable—the extent to which the high-information treatment conveys information on the basis of which employers statistically discriminate.

Aside from the focus on statistical discrimination, a few studies try to test Becker's model of customer discrimination by comparing callback rates across jobs with different degrees of customer contact. (I am not aware of attempts to test for employee discrimination, which would require information on the composition of the tested employers' workforces.) Finally, a couple of recent studies, discussed below, use creative means to try to examine whether implicit discrimination accounts for findings of discrimination in AC studies.

7.5 *Studies of Race, Ethnicity, and Sex Discrimination*

7.5.1 *Summary of Findings*

An extensive survey of results from field experiments through 2000 is provided in

Riach and Rich (2002), and Rich (2014) surveys field experiments conducted since then. Both surveys focus on the standard demographic categories. I do not provide a detailed study-by-study discussion of the extensive evidence on race, ethnic, and sex discrimination that these surveys cover. Rather, I draw out the main lessons from the surveys and focus on issues that may pose challenges to the conclusions from these studies, and hence motivate further research.[56] In subsequent sections, I turn in more depth to some of these issues and to evidence on other groups, especially when that evidence raises new issues that pose interesting areas for additional research. Most of the studies I discuss are summarized in table 3.

Riach and Rich cover eight AC studies of race or ethnic discrimination (which they refer to with the catchall "race") in the United Kingdom, six studies in other European countries, five studies in the United States, and one each in Canada and Australia. Some of these studies report more than one analysis (e.g., for both interviews and job offers, or for different ethnic groups). The key result is that every single comparison in these studies is in the direction of discrimination against the minority group, and most of the estimates of discrimination are statistically significant.[57] The authors conclude that "In view of the number of studies involved and their geographical extent this is compelling evidence of enduring and pervasive racial discrimination in employment" (Riach and Rich 2002, p. F499).

The survey of more recent studies in Rich (2014)—covering a larger number of studies and more countries—also documents near-uniform evidence of hiring discrimination against racial and ethnic minorities, with only a handful of exceptions. One notable finding across a number of studies (in Australia, Belgium, France, Germany, Italy, and Sweden) is the consistent evidence of labor market discrimination against applicants of Middle Eastern origin (often immigrants).[58]

The findings on studies of sex discrimination are more ambiguous. Riach and Rich (2002) summarize results for five studies—three in the United States, and one each in Australia and Austria. An interesting dimension of these studies is that some of them distinguish between results for more traditionally male- versus female-dominated fields or sectors. These studies appear to provide clear evidence of discrimination against

[56] Riach and Rich (2002) also provide a detailed response to many points raised in the Heckman and Siegelman (1993) chapter critiquing the Urban Institute audit studies. (See also Pager 2007.)

[57] The significance tests are based on the net discrimination measure that excludes from the denominator tests with no positive responses. This has no effect on the sign of the estimated discrimination, and an ambiguous effect on statistical significance (increasing the estimate by using a smaller denominator, but reducing the sample size).

[58] There is no clear way to explicitly signal religion in these studies. However, a recent study by Weichselbaumer (2015a) studies discrimination against Turkish women in Germany based on whether they wear a headscarf, which is likely to signal a religious Muslim. Compared to women with German names, and women with Turkish names who do not wear a headscarf (in the photos included), Turkish women with a headscarf experienced much lower callback rates (roughly three times the penalty for a Turkish name).

Two correspondence studies in India—where apparently caste and religion can be signaled by name—find mixed evidence. In software jobs, Banerjee et al. (2009) find no evidence of discrimination against lower castes, or between Hindu and Muslim upper-caste applicants. But in call centers they find discrimination against lower castes, which they suggest may be attributable to employers wanting workers with better English, American telephone etiquette, etc., on the basis of which they statistically discriminate against lower-caste applicants. Siddique (2011) finds some evidence of discrimination against lower castes in white-collar jobs, concentrated on female applicants in front-office or administrative jobs. She also finds some evidence that caste-related differences in callback rates vary with the sex and religion of recruiters (inferred from the name of the contact person listed in the ad, who may not be the decision maker), and suggests that since expectations regarding the different groups of applicants should be the same across recruiters, variation in callback rates with recruiter characteristics is more likely to reflect taste discrimination.

TABLE 3
SELECTED RESULTS FROM FIELD EXPERIMENTS ON LABOR MARKET DISCRIMINATION

Study	Method	Summary/Results
Race and ethnicity		
Riach and Rich (2002)	Survey of existing studies through 2000.	*Consistent evidence of hiring discrimination against racial and ethnic minorities in many countries.* Every comparison points to discrimination against the minority group, most statistically significant.
Rich (2014)	Survey of existing studies from 2000 through around 2012.	*Consistent evidence of hiring discrimination against racial and ethnic minorities in many countries.* Across many studies, near-uniform evidence of hiring discrimination against racial and ethnic minorities, with few exceptions. Many studies for European countries document evidence of discrimination against applicants of Middle Eastern origin (often immigrants).
Carlsson and Rooth (2012)	Combine data from correspondence study in Sweden with survey information on attitudes toward immigrants, which vary at the municipal level.	*Callback rate for Middle Eastern minorities was lower in municipalities where respondents had more discriminatory attitudes, interpreted as consistent with a role for taste discrimination*. However, the attitudinal question could reflect population differences and hence the results could reflect statistical discrimination.
Zschirnt and Ruedin (2016)	Meta-analysis of studies of race and ethnicity from OECD countries, for 1990–2015.	*Consistent evidence of hiring discrimination against racial and ethnic minorities in many countries. Mixed evidence regarding taste versus statistical discrimination, and interpretation not always clear*. Near-uniformity of findings of hiring discrimination against ethnic minorities. Discrimination somewhat lower for second-generation than first-generation immigrants, but not significantly. Discrimination highest for Arabs/Middle Eastern origin, followed by Indians, Pakistanis, Bangladeshis, and Chinese, and then Turks, interpreted as taste discrimination associated with the most distant groups. Discrimination lower in German-speaking countries, where job applications are much more detailed, which the authors interpret as consistent, in part, with statistical discrimination.
Oreopoulos (2011)	Correspondence study of immigrants from many countries in Canada, across multiple jobs.	*Evidence of hiring discrimination against immigrants. Evidence often not consistent with statistical discrimination.*
Rooth (2010)	Incorporation of responses to IAT and explicit discrimination measures into data from correspondence studies of discrimination against Arab-Muslim men in Sweden.	*Implicit discrimination may account for ethnic discrimination.* Recruiters with higher IAT (more discriminatory) less likely to call back Arabs. Responses to explicit measures do not explain relationship, nor do explicit measures significantly predict callbacks. Evidence interpreted as showing implicit discrimination, but because explicit measures do not condition on applicant characteristics, they may not reflect discriminatory attitudes correctly.
Bartoš et al. (2016)	Test for differences in "attention" in employer evaluation of job applicants by measuring parts of resumes examined by employers or efforts toward getting resume.	*"Attention discrimination" may disadvantage minority applicants, causing employers to spend less time evaluating them, and reinforcing emphasis on group averages of disadvantaged groups*. In the Czech study, employers call back minorities at lower rates, and pay less attention to the minority resumes, in terms of both opening resumes and looking at more information—although these differences are generally not significant. In the German study employers exert less effort to obtain minority resumes.

(*Continued*)

TABLE 3
Selected Results from Field Experiments on Labor Market Discrimination (*Continued*)

Study	Method	Summary/Results
Neumark (2012), Carlsson, Fumarco, and Rooth (2013), Baert et al. (2015)	Application of heteroskedastic probit method from Neumark (2012) to existing data or new data.	*Allowing for different variances of unobservables tends to increase estimated discrimination against racial/ethnic minorities. No indication that existing studies badly overstate discrimination*. In two of the three studies, evidence of larger variance of unobservable for minority group, and larger estimate of discrimination, consistent with a low level of standardization of resumes. Estimates of the relative variances are imprecise, but allowing the difference does not reduce precision of estimate of discrimination.
Sex		
Riach and Rich (2002)	Survey of existing studies through 2000.	*Less clear evidence of general discrimination against women than of hiring discrimination consistent with sex stereotyping of jobs*. Across three US studies, and one for Austria and Australia, evidence points to discrimination against women in male and/or high-status/high-pay jobs, and against men in female jobs. No clear evidence for sex-integrated jobs.
Rich (2014)	Survey of existing studies from 2000 through around 2012.	*Absence of evidence of general discrimination against women in hiring, and more evidence consistent with sex stereotyping of jobs.* In studies for China, England, and France, absence of consistent evidence of discrimination against women, and more consistent evidence of discrimination against women in male-dominated jobs and against men in female-dominated jobs (possibly stronger).
Age		
Bendick, Jackson, and Romero (1997)	Correspondence study for management information systems (men), executive secretaries (women), and writer/editor jobs, 57 vs. 32.	*Strong evidence of discrimination against older workers, and some evidence of statistical discrimination.* Evidence of age discrimination for men and women. Weaker evidence for applications de-emphasizing age and emphasizing youthful qualities.
Bendick, Brown, and Wall (1999)	Audit study for entry-level sales or management jobs, almost all applicants male, 57 vs. 42.	*Strong evidence of discrimination against older male workers.* Evidence of age discrimination, mainly at pre-interview stage, and for job characteristics as well when offers made to both testers.
Lahey (2008)	Correspondence study, general entry-level jobs, 35/45 vs. 50/55/62.	*Strong evidence of discrimination against older female workers.* Evidence of age discrimination, increasing in age. No clear evidence from tests intended to detect statistical discrimination, or employer, employee, or customer discrimination.
Riach and Rich (2006, 2010), Riach (2015)	Correspondence studies for four European countries, focused mainly on jobs as waiters, but also female applicants in retail (in England), 47 vs. 27.	*Strong evidence of discrimination against older males, but not older females.* Clear evidence of age discrimination against older men for waiter jobs in all countries. Evidence of favoritism toward older female applicants in retail (one country only).

(*Continued*)

TABLE 3
SELECTED RESULTS FROM FIELD EXPERIMENTS ON LABOR MARKET DISCRIMINATION (*Continued*)

Study	Method	Summary/Results
Neumark, Burn, and Button (forthcoming)	Large-scale US correspondence study, focused on sales (male and female), administrative assistants (female), and janitors and security (male), 29–31/49-51/64-66.	*Robust evidence of age discrimination against older women. Less clear evidence for men.* Evidence for men not robust to using low-experience vs. commensurate-experience resumes for older men, or correcting for difference in the variance of unobservables.
Farber, Silverman, and von Wachter (2017)	US correspondence study of administrative support job applications for women, 35–37/ 40–42/55–58.	*Evidence of age discrimination against older women.* Significantly lower callback rates for oldest group of applicants.
Obesity and looks		
Rooth (2009)	Correspondence study for Sweden, multiple occupations, photos digitally manipulated to make non-obese applicant look obese.	*Evidence of discrimination against obese men and women.* For both men and women, and most occupations, obesity lowers the callback rate. An attractiveness rating explains the obesity differential for men but not for women, although it is hard to distinguish the two effects. Some indication that obesity discrimination is larger in jobs requiring customer contact (such as sales).
Agerström and Rooth (2011)	Incorporation of responses to IAT and explicit discrimination measures into data from correspondence study of discrimination against obese applicants in Sweden.	*Implicit discrimination may account for obesity discrimination*. Lower callbacks to obese applicants from recruiters with higher IAT (more discriminatory). Responses to explicit measures do not explain this relationship, nor do explicit measures significantly predict callbacks. Evidence interpreted as showing implicit discrimination, but explicit measures do not condition on applicant characteristics.
López Bóo, Rossi, and Urzúa (2013)	Correspondence study in Argentina, with digitally manipulated photos, across multiple occupations.	*Ambiguous evidence of discrimination based on looks.* Callback rates for attractive candidates one-third higher, similar for men and women. But no difference when photos are subjectively ranked.
Ruffle and Shtudiner (2015)	Correspondence study in Israel, based on subjectively rated photos, across multiple jobs.	*Discrimination based on looks for men, more so in jobs requiring experience. No evidence that customer contact matters.*
Galarza and Yamada (2014)	Correspondence study in Peru, based on subjectively rated photos, across multiple occupations.	*Evidence of discrimination based on looks and race, but lookism appears to explain racial difference.* Among unskilled jobs, looks matter only in jobs with customer contact.

(*Continued*)

TABLE 3
Selected Results from Field Experiments on Labor Market Discrimination (*Continued*)

Study	Method	Summary/Results
Women with family responsibilities		
Duguet and Petit (2005), Petit (2007)	Correspondence study for financial sector in France, men and women of different ages and family status associated with likely future children.	*Evidence of discrimination against young women based on expected future childbearing, in jobs where turnover may matter more.* No significant differences in callback rates by sex. For higher-skill jobs with higher turnover costs there is evidence of lower callbacks for women aged 25 and single.
Correll, Benard, and Paik (2007)	US correspondence study designed to match lab experiment in same paper.	*Evidence of discrimination against women with childrearing responsibilities.* No significant difference in callback rates for men dependent on parental status, but a callback rate for childless women 1.8 times that for mothers.
Criminal background		
Pager (2003)	Audit study in Milwaukee, black and white tester pairs, broad array of low-skill jobs.	*Clear evidence of differential treatment, which may not reflect discrimination.* For whites, callback rates were lower by 50 percent for those with criminal backgrounds; among blacks, callback rates were lower by 71 percent.
Baert and Verhofstadt (2015)	Correspondence study in Belgium, across multiple occupations.	*Clear evidence of differential treatment for young men with juvenile delinquent past, which may not reflect discrimination.* The callback rate for male applicants with a criminal background was 42 percent lower.
Sexual orientation		
Weichselbaumer (2003)	Correspondence study in Austria, secretary and accountant jobs.	*Evidence of hiring discrimination against lesbians.* Lesbians receive substantially fewer callbacks, with no difference depending on masculinity or femininity. Unlikely to reflect statistical discrimination.
Drydakis (2009)	Correspondence study in Greece, office, industrial, restaurant, and retail jobs.	*Evidence of hiring discrimination against gays.* Substantially lower callback rates for gays, although no difference in wages of jobs for which both gays and straights invited to interview.
Drydakis (2011)	Correspondence study in Greece, office, industrial, restaurant, and retail jobs.	*Evidence of hiring discrimination and consistent with wage discrimination against lesbians.*
Tilcsik (2011)	Correspondence study in United States across many jobs sought by new college graduates.	*Evidence of hiring discrimination against gays, perhaps mediated by antidiscrimination laws.* Significantly lower callbacks for gay applicants, but only in some states. Multivariate analysis suggests antidiscrimination laws reduce callback differentials, and that gays were less likely to receive callbacks from employers with ads referencing stereotypically male traits.
Bailey, Wallace, and Wright (2013)	Correspondence study in United States of job listings on *CareerBuilder.com* ®.	*No evidence of hiring discrimination against gays or lesbians.*

(*Continued*)

TABLE 3
SELECTED RESULTS FROM FIELD EXPERIMENTS ON LABOR MARKET DISCRIMINATION (*Continued*)

Study	Method	Summary/Results
Mishel (2016)	Correspondence study in United States for administrative jobs	*Evidence of hiring discrimination against lesbian/queer women.*
Drydakis (2014)	Correspondence study in Cyprus, with variation in information provided, office, industrial, restaurant, and retail jobs.	*Evidence of hiring discrimination and consistent with wage discrimination against gays and lesbians.* Lower callback rates for gays and lesbians, and lower potential wage offers. No evidence that more information affects measured discrimination.
Disability		
Ravaud, Madiot, and Ville (1992)	Correspondence study in France, broad array of employers, paraplegics.	*Evidence consistent with hiring discrimination against disabled.* Lower callback rates for the disabled, more so among less-qualified applicants.
Baert (2016)	Correspondence study in Belgium, for blind, deaf, or autistic applicants, also varying based on eligibility of disabled for wage subsidy.	*Evidence consistent with hiring discrimination against disabled.* Disabled candidates about half as likely to receive a positive callback. Results reported as not sensitive to correction for differences in the variances of unobservables.
Ameri et al. (2018)	Correspondence study in the United States, applicants with spinal cord injuries or Asperger's syndrome to accounting jobs.	*Evidence consistent with hiring discrimination against disabled, and with an effect of antidiscrimination provisions in reducing this discrimination.* Significantly lower callbacks for disabled applicants. Gap larger for more-skilled applicants, and small and not statistically significant for less-experienced applicants. A variety of analyses suggest that discrimination is concentrated among employers below the ADA cutoff of 15 employees.
Evidence from anonymized application experiments		
Krause, Rinne, and Zimmerman (2012)	Anonymize applications for selection for interview at European research institute.	*Possible evidence of discrimination in favor of women, or anonymization does not increase interviewing of women.* Female applicants were more likely to be invited for interviews in the non-anonymous sample, whereas this advantage was erased in the anonymous sample, although this could be due to hiring committee's knowledge of experiment.
Behaghel, Crépon, and Le Barbanchon (2015)	Anonymize applications for selection for interview in France.	*Possible evidence of discrimination in favor of minorities, or anonymization does not increase interviewing of minorities*. Minority applicants were less likely to be interviewed and hired when applications were anonymized. Possibly attributable to inability to downweight negative signals for minorities, or to engage in preferential hiring of minorities.

women in male-dominated fields, and also in higher-pay or more senior jobs within a sector (e.g., high-price and higher-pay restaurants in the United States in my 1996 study), but also discrimination against men in female-dominated jobs. Moreover, some studies fail to find evidence of sex discrimination in more sex-integrated jobs.

The additional studies surveyed in Rich (2014) confirm both findings. Indeed, Rich suggests that some of the strongest evidence of discrimination arises for men applying to typically female jobs (although this is largely based on net discrimination measures computed relative to tests for which one or both testers received a positive response).[59] Emphasizing the absence of consistent evidence of discrimination against women, in a study of China, Zhou, Zhang, and Song (2013) found statistically significant evidence of discrimination against men in most jobs covered, including many jobs that would not be viewed as typically female.

A recent meta-analysis by Zschirnt and Ruedin (2016) also provides summary measures, in a more consistent form, of correspondence studies of racial and ethnic discrimination in OECD countries. The overlap with Riach and Rich (2002) and Rich (2014) is substantial, and the simple summary statistics therefore, not surprisingly, convey the same near-uniformity of the findings of hiring discrimination against ethnic minorities.[60] What is unique in this study, however, is the attempt to quantify (via meta-analysis) how the estimates of discrimination in these studies vary with a number of study characteristics, to attempt to shed light on hypotheses about what drives relatively stronger or weaker evidence of this discrimination, and to try to understand the nature of discrimination. Most of this part of the analysis pertains to trying to test for statistical versus taste discrimination, which I take up below. However, another interesting finding the authors report is that they find no association between GDP growth or unemployment rates and measured discrimination, in contrast to the hypothesis that employers are more likely to engage in discrimination in slack labor markets (e.g., Biddle and Hamermesh 2013).[61]

7.5.2 *Testing for the Nature of Discrimination by Race, Ethnicity, and Sex*

Based on the consistent evidence of hiring discrimination across so many groups, Riach and Rich (2002) suggest that it is unlikely that statistical discrimination underlies most of the findings:

> Given the very diverse cultural, social, religious and educational backgrounds of the various Asian, Arab and African-descendant groups who have encountered the employment

[59] And based on data from correspondence studies of sex discrimination in Sweden, Carlsson (2011) concludes that the relationship between gender bias in callbacks and the female share in the occupation is quite weak. He also finds that the gender of the recruiter and the proportion female at the firm (based on matching to firm-level data) does not affect the relative treatment of female candidates, which he interprets as ruling out "in-group favoritism" (p. 90).

[60] Similarly, a very recent meta-analysis of US studies on discrimination against blacks and Hispanics finds that the evidence almost always points to discrimination against minorities (Quillian et al. 2017).

[61] In contrast, Farber, Silverman, and von Wachter (2017) report less evidence of age discrimination among employers with higher overall callback rates, which they interpret as "employers with a high demand for workers become less selective in deciding whether or not to call back" (p. 6). Baert et al. (2015), in a study of discrimination against Turks in Belgium, find less ethnic discrimination in what they term "bottleneck" occupations, where vacancies take longer to fill (and which tend to be higher skilled). This does not necessarily speak to changes in discrimination over the business cycle. It could be related, for example, to the importance of a better match in some occupations than others, generating both longer-duration vacancies and less tendency by employers to discriminate—paralleling the evidence on Chinese job boards indicating that, for higher-skilled jobs, employers are less likely to express preferences based on worker demographics (Kuhn and Shen 2013).

> discrimination in these studies, it would be a superficial generalization to hold that the recorded disinclination to hire them arises because the various "indigenous" white populations have, on average, superior employment characteristics. The common characteristic of Asians, Arabs and African-descendant groups is that they are not white …. White immigrant groups (for example, Italians, Greeks) encountered discrimination in some studies, but never at a level comparable to non-whites (pp. F499, F503).

Of course, we cannot decisively rule out statistical discrimination based on this observation. Nonetheless, the argument is somewhat compelling.

Similarly, the findings on sex discrimination are hard to square with statistical discrimination. The usual argument is that employers engage in statistical discrimination against women based on a higher likelihood of leaving the firm, taking on family responsibilities, etc. It is not clear, then, why the evidence from AC studies points to discrimination against men in typically-female jobs, and vice versa, rather than discrimination against women generally, although it is in principle possible that these stereotypical female behaviors are for some reason more compatible with female jobs, and similarly for male behaviors and characteristics and male jobs. But that would seem to lead to a rather tortured argument about statistical discrimination, and in this case, coupled with the evidence on race and ethnicity, Occam's razor might well argue for a simpler taste-based discrimination model—adapted in the case of hiring discrimination by sex to have more to do with violating stereotypes or norms than disutility from hiring women.

Some AC studies try to test for statistical versus taste discrimination more directly, by manipulating the amount of information provided. Rich (2014) concludes that the evidence is inconsistent across studies, with some finding that including more information eliminates group differences (e.g., Kaas and Manger 2012), whereas many do not. As explained above, however, these tests may not be informative, and the same hypothesis can generate different results in different studies because of the relation between study design, what employers know, and what information is provided.

Other studies report evidence that authors argue is not consistent with either model. For example, Bertrand and Mullainathan (2004) suggest that their evidence of similar magnitudes of race discrimination across occupations is hard to reconcile with statistical discrimination because factors such as the importance of unobservable skills and the ease of observing relevant characteristics should vary across occupations. But they also note results that are hard to reconcile with some taste-based models—for example, the similarity of results across occupations with different degrees of customer or coworker contact. Rich (2014) claims that, across studies, measured discrimination does not vary with customer contact, although no formal analysis is presented.

Carlsson and Rooth (2012) combine data from their correspondence study in Sweden with municipal-level survey information on attitudes toward immigrants. In evidence they interpret as consistent with taste discrimination, they find that the callback rate for Middle Eastern minorities was lower in municipalities where respondents had more discriminatory attitudes. However, the attitudinal question refers to the value of immigrants in contributing to the Swedish population. It is not entirely clear what this means, but is possible that the productivity of the immigrant population, relative to nonimmigrants, differs across municipalities, in which case the reported differences in attitudes could reflect these population averages, and the relationship with callback rates could reflect statistical discrimination. Attitudinal questions focused more on discriminatory tastes or preferences would likely be more convincing.

Zschirnt and Ruedin's (2016) meta-analysis focuses on statistical versus taste discrimination, presenting three types of evidence. First, they suggest that statistical discrimination should be lower for second-generation than first-generation immigrants because more information is available about the former; for example, more of the qualifications (like schooling) are local. They find some evidence of lower discrimination against second-generation immigrants, although the estimates point to substantial discrimination for both groups and the difference is not large. In addition, there are other reasons measured discrimination may change across immigrant generations, such as increased proficiency in the native language, which may not be adequately controlled for in a correspondence study. And it is not clear why taste discrimination cannot decline for more assimilated second-generation immigrants.

Zschirnt and Ruedin (2016) suggest that evidence of stronger measured discrimination against "more distant and visible minority groups" (p. 5) would be indicative of taste discrimination. Their analysis suggests the lowest callback rates for Arabs and those of Middle Eastern origin, followed by Indians, Pakistanis, and Bangladeshis as well as Chinese, with Turks having callback rates that are closest to natives, but still lower. Again, though, it is difficult to distinguish taste from statistical discrimination, as group differences or availability of information about these groups may also differ.

Third, the authors note that in German-speaking countries, far more detailed information is included as part of job applications (like diplomas and transcripts), and hence it is less likely that employers need to resort to statistical discrimination. Thus, smaller measured discrimination against minorities in these countries would be consistent with statistical discrimination. Their evidence points to lower measured discrimination for minorities in the German-speaking countries, although the remaining difference in German-speaking countries, despite the extensive information provided, would suggest that taste discrimination remains important. Echoing the discussion above, however, this source of variation is only informative if the studies in each country provide the information employers typically have. Another limitation of this evidence is that it does not condition on the ethnic group, whereas a difference in the German-speaking countries for the same ethnic group, relative to natives, would be more informative about the difference between the two sets of countries owing to differences in the application process. The authors also suggest that much weaker evidence of hiring differences in the public sector is consistent with statistical discrimination, given that the public sector uses more formal and detailed applications. Of course, the public sector may simply engage in less discriminatory behavior as a matter of policy.

Oreopoulos (2011) uses the "information-treatment" test for statistical discrimination against immigrants in a correspondence study in Canada. Some of his analyses are potentially more informative because they implement this test across jobs and skills for which we might expect statistical discrimination to be less important or more important, while the other limitations of this test that I have described are, in a sense, held constant. One approach is varying whether an applicant with a foreign name shows Canadian education and experience, or foreign education and experience. The latter may provide a noisier signal, although the domestic education or experience could have different value in terms of productivity. Oreopoulos also compares results for which relatively more or fewer of the required skills might be inferred from immigrant status. For example, in sales jobs, language might matter a lot, but in computer programming it should not. He does not find consistent evidence

pointing to statistical discrimination from adding information related to important skills. For example, listing English or French fluency increases callback rates for foreign-educated and experienced applicants, but not foreign-named applicants. And he does not find the expected relationship between information about language skills and the skills required on the job, even though some recruiters who responded to a survey said that a foreign name is viewed as a signal of a lack of critical language skills.[62]

Finally, Rooth (2010) tried to test for implicit discrimination in the context of a correspondence study, by going back to the recruiters from his earlier correspondence studies of discrimination against Arab-Muslim men in Sweden and administering the IAT to some of them. The evidence shows that those with a higher IAT (more discriminatory) were less likely to call back Arabs. This result is valuable, as it appears to provide some validation of the IAT in the real world.

Rooth then includes three explicit discrimination measures in the callback model. The evidence indicates that recruiters have explicit discriminatory attitudes. For example, in the data from one experiment, 54 percent say that they prefer hiring a Swedish male to an Arab Muslim male. Rooth finds that these explicit measures do not account for the lower callback rates associated with a higher IAT. He concludes from this evidence that "there are recruiters who implicitly discriminate, but who would not explicitly do so" (p. 529).

This conclusion may not be warranted, however, because the explicit measures are unconditional. For example, the hiring preference measure asks "which group, native Swedes versus Arab-Muslim, they prefer when hiring people," and the performance stereotype questions have alternatives like "Swedish men perform much better at work than Arab-Muslim men" (p. 527). The responses can therefore reflect actual differences in productivity, rather than how a recruiter would treat two candidates with equal qualifications. The unconditional nature of the explicit discrimination measures seems likely to lead to understatement of the effects of these measures, because recruiters who would prefer a Swede unconditionally might not have that preference conditionally (on the characteristics for which the study controls), so the explicit measures may misclassify recruiters with regard to their explicit biases in the conditional experiments. That is, the correlation of the IAT with explicit *conditional* discrimination measures might be much higher. Nonetheless, the idea of combining data from AC studies with the IAT is very creative, and future work could likely make more definitive use of such information.

Finally, in a recent paper, Bartoš et al. (2016) propose a model of statistical discrimination that endogenizes effort devoted to learning about applicants (and hence, in a sense might be viewed as blending statistical and implicit discrimination). They consider costs and benefits of acquiring information about applicants to model how much attention employers will pay to applications from different groups and in different settings. The model predicts that in highly selective markets, employers will pay less attention to applications from groups with lower qualifications. The authors test this prediction in correspondence studies using Czech and German data with native and ethnic minority names. In the Czech study, they measure "attention" by modifying the research design

[62] A different kind of approach is taken in Hedegaard and Tyran (2018). They run an experiment where individuals are given the opportunity to choose from among two coworkers (one of the same ethnic group and one of a different ethnic group). In one treatment, they have information about productivity, and in the other they do not. They find evidence of discrimination in both cases, leading them to conclude that taste-based discrimination plays an important role.

to force employers to click on hyperlinks in emails to see resumes (having seen the name that signals minority status in the email), and then in the resume embedding "learn more" buttons (like clicking on "Experience" to see more about the applicant's job responsibilities). The authors find evidence of discrimination against minorities (Roma and Asians) similar to the many studies already discussed. They also find evidence that employers pay more attention to the nonminority resumes, in terms of both opening resumes and looking at more information—although these differences are generally not significant in their very small sample.[63]

In the German study, a different design is used in which employers have to open a resume, but then an error is generated and they have to take additional effort to try to see the resume. In this case, no information on callbacks is obtained, since the resume is never provided, and the authors obtain information only on information acquisition, for which they find less effort to get resumes of minority (Turkish) applicants.

7.5.3 *Studies Addressing the Heckman Critique*

In Neumark (2012), I applied the method for addressing the Heckman critique to the data from the Bertrand and Mullainathan (2004) study of black-sounding names, exploiting data on skills included on some of their resumes—which do help predict callbacks. The estimated overall marginal effects of race from the heteroskedastic probit model indicate callback rates that are lower for blacks (black-sounding names) by about 2.4–2.5 percentage points, relative to a 9.65 percent callback rate for whites. The estimated variance of the unobservable is larger for blacks (although the difference is not statistically significant). As a result, decomposing the marginal effect—the effect via the level of the latent variable, which captures discrimination as conventionally defined in AC studies—is larger than the marginal effect from the probit estimation, ranging from −5.4 to −8.6 percentage points. The effect of race via the variance of the unobservable, in contrast, is positive, ranging from 2.8 to 6.2 percentage points (although this effect is not statistically significant). The point estimates imply that, in these data, the bias from different variances of the unobservable leads to understatement of race discrimination (presumably reflecting both the higher variance for blacks and standardization of the resumes at a low quality level).

Carlsson, Fumarco, and Rooth (2013) reexamine data from four previous studies of the Swedish labor market by some of the coauthors, each of which includes some form of the data required to implement this method. Their reanalysis does not lead to large changes in the estimates of discrimination, and sometimes the estimated discrimination (against those with Arabic names, and in favor of women) becomes smaller—which contrasts with the findings from the Bertrand and Mullainathan (2004) data. Baert et al. (2015) implemented this method in a study of discrimination against Turkish school-leavers in Belgium, using information on distance from the worker's residence to the workplace and other application characteristics to identify the heteroskedastic probit model, and report that this correction does not alter the conclusions. They find a higher variance of the unobservable for Turks, and a larger estimate of discrimination, consistent with a low level of standardization; again, the difference in variances is not statistically significant.[64]

[63] They test for differences in attention based on selectivity of markets by comparing data for labor and housing markets, arguing that the latter are much less selective.

[64] Baert (2015) does a similar analysis for a field experiment on sex discrimination and finds no effect of this correction on estimates of discrimination.

The estimates from these studies point to understatement of discrimination when the variances of the unobservables are assumed to be equal, and none indicate that ignoring the Heckman critique leads to strong overstatement of discrimination. In contrast, a very recent study (Neumark and Rich forthcoming) implements this technique for five labor market (and four housing market) studies of racial or ethnic discrimination that we identified that include the requisite data and for which we could get the data from the authors. We find that the evidence of race and ethnic discrimination in the labor market studies falls to near zero or becomes statistically insignificant in about half of the estimates reexamined, suggesting that we may need to interpret the overall body of evidence from field experiments on race and ethnic hiring discrimination more cautiously than the other surveys and meta-analyses discussed above.

7.6 *Studies of Age Discrimination*

7.6.1 *Summary of Findings*

A number of studies apply AC methods to age discrimination. In general, this work follows the paradigm used in studies of discrimination against other groups, such as blacks or women, making applicants identical (up to random variation) in all respects except age. However, there is an issue of how to treat potential age-related differences in experience, discussed below.

There are also other issues that may affect the interpretation of studies applying AC methods to the question of age discrimination. For example, employers might interpret older applicants as sufficiently unusual for some jobs that they discount their applications. Although this might still be construed as age discrimination, researchers try to avoid this problem by applying to jobs for which older workers do in fact apply (e.g., Neumark, Burn, and Button forthcoming). Second, employers might negatively evaluate an older worker who is applying for fairly entry-level positions. This would still, however, constitute statistical discrimination, although it is likely less interpretable as taste discrimination. Neumark, Burn, and Button (forthcoming) also discuss this issue, and discount it because evidence of lower callbacks to older applicants is similar for older applicants moving from higher- to lower-skilled jobs near the age of retirement.

The main earlier studies almost uniformly find evidence of age discrimination in hiring.[65] For example, Bendick, Jackson, and Romero's (1997) correspondence study looks at thirty-two and fifty-seven-year-old male and female applicants (signaled by year of graduation). Among applications in which at least one of the two applicants received a positive response, in 43 percent of cases only younger applicants received the positive response, versus 16.5 percent of cases in which the older applicant was favored, for a statistically significant net discrimination measure of 26.5 percentage points.

This study had other interesting features. First, the authors sent applications to a large list of companies that they could identify (without reference to job postings). Drawing on other information sources, they could study how firm characteristics such as firm size or industry affect the outcomes of their tests. For example, they find more evidence of discrimination in the largest firms; this contrasts with other research on discrimination (e.g., Holzer 1998), but that may be because this other research contrasts very small firms that may be exempt from antidiscrimination laws.

A second feature was to vary the extent to which resumes de-emphasized age (by focusing on skills rather than history), high-

[65] In this section, I exclude a few studies that focus on age differences but do not include age ranges covering older workers (in their 50s or 60s).

lighted experience and maturity, or tried to combat age-related stereotypes (by adding to the cover letter a phrase about being "energetic, adaptable to the latest technology, and committed to my career"). Measured discrimination was about half (but still present) for the applications de-emphasizing age or countering ageist stereotypes. The finding on stereotypes might be viewed as evidence of statistical discrimination against older workers based on assuming they are less adept at new technology, or are likely to retire soon. In light of the discussion of the potential limitations of tests for statistical discrimination, this evidence might be viewed as relatively more compelling, since employers likely do care about these things (recall the evidence from Rosen and Jerdee 1977) and would not be able to glean it directly from job applications.

Bendick, Brown, and Wall (1999) perform an audit study meant to provide more evidence on age discrimination than can be obtained from a correspondence study. One goal was to capture more than just whether the callback was positive—such as whether, when both testers received positive responses, one was treated more favorably with regard to getting an interview, an opportunity to demonstrate skills, a job offer, or a job offer with higher compensation. The percentage of tests with a more favorable response for younger applicants (age thirty-two) was 42.2 percent, versus 1 percent for older applicants (age fifty-seven), and the evidence points to favorable treatment of younger workers, among pairs offered jobs, with respect to salary and health insurance benefits.[66]

Another feature of this study is to document differences in outcomes at different stages of the process. Echoing the earlier conclusions from studies of ethnic and race discrimination, about three-quarters of the discriminatory difference in treatment is at the preinterview stage. The study also reports on some assessments of phone calls and interviews by the testers, including attempts to gauge whether employers expressed age-related stereotypes, finding that this was rare. Finally, the study presents information on characteristics of the positions offered, with some indication of less attractive positions offered to older applicants (e.g., lower commissions on sales, part-time work, etc.), although any such conclusions are based on a handful of observations.

Lahey's (2008) correspondence study focuses on women only (for reasons discussed more below). She uses five different ages ranging from thirty-five to sixty-two, and finds consistent evidence of age discrimination, whether using age as a continuous variable or classifying workers as older (fifty/fifty-five/sixty-two) versus younger (thirty-five/forty-five). Lahey tries to address the question of statistical discrimination by including language on older resumes intended to combat adverse stereotypes of older workers (like a statement about willingness to embrace change in the resume), information on an attendance award (to combat expectations of greater absences among older workers), a statement about not needing health insurance (to counter expectations of higher health insurance costs), and information about computer skills (to combat fears of technological obsolescence). There were not clear indications that resume features designed to counter negative age-related stereotypes consistently helped older applicants.

Lahey also tries to assess the importance of employer, employee, or customer discrimination. She tests for employer discrimination by assuming that a firm with a human resources department would exhibit less taste-based discrimination (although it is unclear why the department would not act at the owner's discretion), but finds evidence (not significant) suggesting a negative interaction with applicant age. She tests for employee discrim-

[66] Details in the paper are scant, including statistical tests for each dimension of job offers.

ination by asking whether the age composition of the workforce in the employer's Public Use Microdata Area (PUMA) is related to variation in outcomes with age, but finds no relationship (acknowledging that this is a noisy measure, and hence imprecise test). Finally, she does a related test for customer discrimination by asking whether the age composition of the population is more important in occupations with more customer contact; again, she finds no evidence of a relationship.

Three closely related studies by Riach (2015) and Riach and Rich (2006 and 2010) investigate age discrimination in applying for jobs as waiters (and, in one case, as retail managers). One paper covers England, one covers France, and Riach (2015) covers four countries, adding Spain and Germany. In all instances but one (applicants for jobs as retail managers in Riach and Rich 2010, who were female only), there is strong evidence of discrimination against older applicants. This one instance is the only one I am aware of in the research literature that finds evidence of discrimination against younger workers. However, for correspondence studies, the samples in all of these studies are quite small; there are only 300 retail manager applications, and the net discrimination measure is based on 24 applicants.

7.6.2 *Specific Issues in Using Correspondence Studies to Test for Age Discrimination*

Strict application of the paradigm of AC methods to studying age discrimination requires giving older and younger applicants the same low level of experience (commensurate with the young applicants' ages), because clearly a young applicant cannot have the experience of a long-employed older worker. This could make older applicants look unusually unqualified relative to the older applicants employers usually see, which could generate a bias in AC studies toward finding evidence of age discrimination. Researchers have addressed this problem in different ways, but they may not have eliminated this bias.

Bendick, Jackson, and Romero (1997) had older and younger applicants report ten years of experience on their resumes. To account for the gap in older applicants' job histories, the resumes indicated they had been out of the labor force raising children (for the female executive secretary applications), or working as a high school teacher (for the male or mixed applications). Bendick, Brown, and Wall (1999) are vaguer, but note that all applicants reported several years of experience in an occupation related to the job to which they are applying, and older applicants also reported additional years of experience in an unrelated field. Lahey (2008) includes only ten-year jobs histories for all applicants, although she focuses on women, for whom time out of the labor force may be less of a negative signal than for men. Still, across these studies it is a matter of speculation whether experience in unrelated fields or time out of the labor force negatively affected employers' assessments of older applicants, generating spurious evidence of age discrimination.

Neumark, Burn, and Button (forthcoming) argue that the more relevant policy and legal question is relative hiring of older workers and younger workers each with experience commensurate with their age. Riach and Rich (2006, 2010) also criticize the approach of using unrealistic resumes for older workers, and in their studies give their applicants experience commensurate with their age. Most of their findings (with the one exception noted above) still find strong evidence of age discrimination, suggesting that low experience does not drive the finding in other studies. However, in their sample resumes, the description of experience is quite cursory, and not a full job history.[67] Results

[67] In the 2006 paper, the resumes simply say that the person has worked as a server in restaurants since about age twenty. In the 2010 paper, one resume lists three jobs

could differ with full job histories for older workers showing experience commensurate with their age.

Neumark, Burn, and Buttton (forthcoming) provide a more comprehensive examination of this question in a large-scale correspondence study of age discrimination, drawing on a comprehensive database of resumes to construct different kinds of older worker resumes showing either experience that matched that of younger applicants, or experience commensurate with age, to see whether there was a difference. For women, perhaps consistent with Lahey's (2008) conjecture, the outcomes relative to younger applicants did not depend on whether older applicants' resumes showed matched or commensurate experience. For men, however, in the one of two occupations in which there was some evidence of age discrimination, the results were fully driven by the low-experience resumes for older workers. Thus, there is some evidence that using low-experience resumes in age discrimination AC studies is problematic, at least for men (for whom, for older cohorts, there may be more of an expectation of a fairly continuous work history).

The second issue emphasized in this study is differences in the variances of the unobservables. Unequal variances of the unobservable may be particularly plausible in comparing older and younger workers. The human capital model predicts that earnings become more dispersed with age as workers invest differentially in human capital and these differences accumulate (Mincer 1974). Because this variation is unlikely to be conveyed on the resumes used in correspondence studies, it could imply a larger variance of unobservables for older versus younger applicants. The correspondence study was designed to include multiple skill variables that could shift callback probabilities, and to then use these variables, in a heteroskedastic probit model, to estimate the relative variances of the unobservables and obtain an unbiased estimate of age discrimination. Paralleling the results on experience, the findings for women were robust to doing this; nothing diminished the rather strong evidence of age discrimination against older women. For men, however, the evidence was not robust to this correction.

The findings in Neumark, Burn, and Button (forthcoming) provide decisive evidence of age discrimination against older women. A recent US correspondence study by Farber, Silverman, and von Wachter (2017) provides corroborating evidence. The study focuses more on the effect of unemployment duration than on age discrimination, but finds evidence of lower callback rates for women aged 55–58 (compared to 35–37 and 40–42) who apply to administrative support jobs (one of the jobs in Neumark, Burn, and Button forthcoming). In contrast to the earlier studies, however, the evidence in this study raises doubts about the conclusion that older men suffer much age discrimination; certainly the finding is not robust. One possible reason older women may experience more discrimination than older men is because physical appearance matters more for women (Jackson 1992), and age detracts more from physical appearance for women than for men (Berman, O'Nan, and Floyd 1981)—consistent with the shift in relative preference away from women with age that Kuhn and Shen (2013) find.

7.7 *Studies of Discrimination Against Other Groups*

AC studies have been conducted for many other groups of workers—some protected by antidiscrimination laws (e.g., the disabled), and some not (e.g., the less attractive). There is typically only a small number of studies for each such group. I review some of these studies briefly, focusing on some of their

held since age seventeen, rising to Senior Waiter, and the second has a paragraph description of the career since leaving school, again with rising responsibility.

interesting features, especially those that might prompt additional research.

7.7.1 *Looks and Obesity*

I discuss experimental research on looks and obesity together. Although there are some observable outcomes associated with obesity that employers may care about and may be health related, it is plausible that negative visual appearance is a common factor in both of these dimensions. Indeed, Rooth's (2009) correspondence study of obesity in Sweden explicitly combines the two. Rooth signals obesity by digitally manipulating photographs—which are often included in job applications in Sweden—to change a non-obese applicant to obese. For men and women, obesity lowers the callback rate by about seven to eight percentage points, and the sign is in this direction for all occupations but nurses. The study also incorporates attractiveness ratings of the photographs, including the digitally altered ones. Attractiveness is rated lower for the obese applicants, and pooling across occupations, the attractiveness rating explains the obesity differential for men, but not for women.[68] However, the two measures are highly correlated, so it is hard to distinguish the effects of looks and obesity. Moreover, across occupations, the relative magnitudes of the effects of the two characteristics of the photos are not robust. The obesity penalty is larger in some jobs requiring customer contact (such as sales), which may reflect customer discrimination. But the occupation-specific estimates are imprecise, and the pattern is not consistent (e.g., a large penalty for female but not male accountant applicants).

Focusing just on obesity, Agerström and Rooth (2011) follow Rooth (2010), testing for implicit biases against the obese by administering the IAT to some recruiters from Rooth's (2009) correspondence study of obesity discrimination and asking about explicit discrimination and obesity-related stereotypes regarding overall work performance. The results are similar to the Arab-Muslim discrimination study. Recruiters with a higher IAT (more discriminatory toward the obese) were less likely to call back obese applicants, controlling for the explicit measures did not change this result, and the explicit measures were not significantly associated with callbacks. However, the previous criticisms apply here as well.

Two studies focus on looks in isolation. López Bóo, Rossi, and Urzúa (2013) study Argentina. They use photographs that are digitally manipulated to alter what they say are validated measures of facial beauty based on horizontal distance between the eyes relative to a face's width, and vertical distance between the eyes and mouth relative to a face's length. They find that callback rates for the attractive candidates are about one third higher, and the effect is similar for men and women. However, when they have students rank the photos, they get high correlations of rankings with the distance measures, but the subjective rankings do not predict callbacks. This raises the question of what is being manipulated in the experiment, since we might expect student ratings of looks to line up to some extent with employers' perceptions.

Ruffle and Shtudiner (2015) do a correspondence study of lookism in Israel, where including a photo is discretionary. They use subjective ratings of attractiveness by a panel of working people and find higher callback rates for more attractive men, but not women. They also send out applications with no photos, and among women, these get the highest callback rate, whereas for men

[68] Rooth used a professional firm to do the photo manipulation, and reports than none of the student evaluators he used suggested that the photos looked odd or manipulated. The López Bóo, Rossi, and Urzúa (2013) study (discussed below) reproduces the original and altered photos, so the reader can judge.

the no-photo applications only get higher callback rates than the non-attractive applicants.[69] These kinds of results raise questions about which real-world applicants actually send out photos, and hence about the external validity of the findings in countries where including a photo is discretionary.[70] As in other studies, the authors test for differences based on interaction with customers. For neither men nor women does the treatment based on looks vary with customer contact.

Finally, Galarza and Yamada (2014) conduct a correspondence study in Peru of discrimination based on looks and race (signaled via photos and names). The photos are manipulated to be similar in terms of clothing, background, etc., and are rated by a group of professionals from various fields (they sound like academics). They include both men and women. Looking only at race (indigenous origins), the authors find strong evidence of racial discrimination in all types of jobs. However, white applicants are rated as better looking than indigenous ones, raising the question of whether the racial discrimination in callbacks reflects lookism or something else. The study finds a significant beauty "premium" in callback rates, with looks explaining about half the racial gap in callbacks. In professional jobs, there is only a difference associated with looks under this approach, and no racial difference. In technical and unskilled jobs, neither appears to matter when both are included in the model. Among the unskilled jobs (in results the authors only describe), looks seem to matter for those with customer contact (sales and marketing, and restaurants), but not in non-contact jobs (e.g., call centers and drivers), consistent with customer discrimination that presumably makes good-looking employees more productive in such jobs. This study, like Rooth's (2009), raises the question of how to interpret the effects of looks, and of other physical characteristics that seem to be associated with looks.

7.7.2 *Women with Family Responsibilities*

Statistical discrimination against women may stem from employers' expectations that they will leave their job, perhaps to have or care for children. Duguet and Petit (2005) and Petit (2007) test this hypothesis in a correspondence study of jobs in the financial sector in France.[71] They do not use the information-treatment approach, presumably because it is hard to imagine a credible direct signal of future turnover or childbearing intentions. Instead, they use groups with different likelihoods of future childbearing or childrearing responsibilities. In particular, they have three types of both male and female applicants: age twenty-five, single, and childless; age thirty-seven, single, and childless; and age thirty-seven, married, with three children. Women in the first group are those for whom employers are most likely to expect future childbearing. Consistent with past studies, they do not find significant overall differences in callback rates by sex. Nor, for all jobs combined, do they find differences for the three kinds of applicants. They do find, however, that for the jobs with higher qualifications or that offer training—jobs in which turnover costs may be higher—the women who are twenty-five, single, and childless receive fewer callbacks than comparable men. In contrast, for the thirty-seven-year-old

[69] The authors present some evidence that they interpret as suggesting that the disadvantage or lack of advantage for attractive female applicants in their data may stem from jealousy or envy among females who evaluate job applications.

[70] The authors report some results from asking their subjective raters of looks whether they would attach the picture to an application if it was their picture, and ask employers which types of applicants tend to attach a picture to their resume and what message applicants convey by including a photo.

[71] The data are the same in the two papers, and the analysis and results are nearly the same; I describe the results from the first paper.

applicants there is either no sex difference or, in one case, a difference favoring women. This evidence is consistent with statistical discrimination against women with regard to future childbearing.

In the Correll, Benard, and Paik (2007) study discussed above, the evidence from the lab experiment is coupled with evidence from a correspondence study. The design of the correspondence study was similar to the lab experiment, with a male pair and a female pair (signaled by names) in each of which one was a parent and one was not, and applications were made to marketing positions; marital status was not indicated. The results indicated no significant difference in callback rates for men dependent on parental status, but a callback rate for childless women 1.8 times that for mothers. Given that parental status was signaled as being a Parent–Teacher Association Coordinator, and applicants had about seven years of experience, the applicants were reasonably interpreted as having young children.

7.7.3 *Criminals/Criminal Records*

Perhaps the least surprising result among field experiments on labor market discrimination is that applicants with criminal backgrounds get fewer callbacks. Pager's (2003) audit study of black and white applicants in Milwaukee finds that, among whites, callback rates were lower by 50 percent for those with criminal backgrounds (a felony drug conviction and eighteen months served), and among blacks, callback rates were lower by 71 percent. Unusually, the study uses direct employer contacts, so that race could be signaled directly, like in an audit study; but it did not have the testers go beyond the first contact, so only callbacks are captured.[72] Strikingly, perhaps, the callback rate for the white tester with a criminal background record is higher (not significantly) than that for the black tester without a criminal background. However, there were only four testers, paired by race, so the comparisons by race do not come from a within-pair design.[73]

Baert and Verhofstadt (2015) carried out a correspondence test in Belgium, using as applicants male school leavers applying to a number of occupations, some of whom indicated a period of juvenile delinquency (spending time in a detention center). The applicants indicating this history also reported having a certificate of good conduct. The callback rate for control applicants who did not report a criminal background was 29 percent (four percentage points) higher. The difference is more pronounced for males, and small and statistically insignificant for females.

Of course, lower callbacks for those with a criminal background should not necessarily be viewed as discrimination, because criminal background signals relevant information that employers not only may care about but perhaps should care about, given that they can be held liable for an employee's actions and accused of negligence in hiring, which a background check does not necessarily protect against (e.g., *Ponticas v. K.M.S. Investments* 1983).[74] In neither country covered by

[72] Like in age AC studies, there is an issue of missing labor market experience, attributable to the eighteen months spent in prison. Pager addresses this by having the testers show some work experience in prison, and having those without criminal backgrounds graduate later. Still, there could be an experience differential that matters for the outcomes, although my sense is this is unlikely to explain much of the difference associated with criminal background.

[73] Pager, Bonikowski, and Western (2009) extend this study (doing it in New York) to include Latino applicants as well, using alternative assignments of testers that permit matched-pair comparisons between nonminority ex-offenders and minorities without a criminal background. The white ex-offenders have modestly higher (though not statistically significant) callback rates than non-offender blacks and Latinos.

[74] Despite this, Baert and Verhofstadt (2015) strongly interpret their results as reflecting discrimination (e.g., p. 1070). A recent experiment (Agan and Starr 2017) estimates the effect of state policies that restrict employers

these studies is discrimination against those with a criminal background explicitly illegal, although some US states are trying to increase employment protections for workers with criminal backgrounds (Pager 2003).[75] Nonetheless, the results do emphasize the challenge of reintegrating those with criminal backgrounds into the labor market.

7.7.4 *Gays and Lesbians*

The research literature using AC methods to study discrimination against gays and lesbians is, perhaps not surprisingly, more recent. Perhaps reflecting this, the studies in this area present some interesting and unusual features, some of which—such as trying to learn about wage offers in the context of these studies—could perhaps be fruitfully emulated and extended in AC studies more generally.

Weichselbaumer (2003) conducted a large-scale correspondence study of discrimination against lesbians in Austria. Her analysis is motivated in part by the evidence of a wage penalty for gays but a wage premium for lesbians, relative to same-sex heterosexuals. Weichselbaumer discusses a number of hypotheses that might explain why lesbians earn a wage premium despite evidence that, like gays, they self-report discrimination in the workplace, and that surveys (at the time of this research) indicated negative attitudes toward both gays and lesbians, including in the workplace. First, the wage premium may be confined to more "masculine" lesbians, if this masculinity is valued in the workplace. Second, the adverse effects of discrimination may not be reflected in wages because few lesbians are "out" at work, whereas a correspondence study that manipulates sexual orientation would reveal discrimination. Third, lesbians may specialize less in home production, including having fewer children, boosting wages despite discrimination.

The study submitted applications for secretarial and accounting jobs, using the detailed applications (including photographs) typical in German-speaking labor markets. Masculinity or femininity is conveyed via the photos, and the study used a group of student raters to confirm that the photos conform to these stereotypes without influencing overall ratings of desirability. The stereotypes were also reinforced via hobbies and past experience on the resumes, although it is possible that these were more strongly correlated with productivity/desirability. Sexual orientation is signaled via work for a gay rights organization, balanced in the heterosexual sample by work for other organizations.[76] The study successively sends

asking about criminal background histories on online job application sites ("ban-the-box" policies), based on callbacks to fictitious applicants to companies—some of which asked this information, and some of which did not—before and after the policy change. The evidence indicates that criminal records are a significant barrier, but also that "banning the box" increased the white versus black advantage in callback rates, presumably because it encouraged statistical discrimination against blacks who could no longer signal at this initial stage their absence of a criminal record.

[75] The EEOC does warn that discrimination on the basis of criminal background can lead to racial or ethnic discrimination, given high incarceration rates for blacks and Hispanics (http://www.eeoc.gov/laws/guidance/arrest_conviction.cfm#I, viewed January 18, 2016).

[76] Baert's (2014) correspondence study of discrimination against lesbians in Belgium signals sexual orientation by having all applicants married, with the lesbian applicants listing a female spouse (based on first name), versus listing no spouse. Baert suggests that other studies' use of membership or participation in a gay rights group may signal activism or even radicalism. However, the other studies usually assigned a similar kind of activism (e.g., in an environmental group) to the controls, so it is not clear why this method should accentuate discrimination against lesbians. Weichselbaumer (2015b) compares results using these two approaches (in Germany, using lesbians in "registered partnerships" to compare with married heterosexual women), and finds little difference in the callback gap between single straight and lesbian women signaling sexual orientation via volunteer activities, and married straight and registered-partner lesbian women signaling sexual orientation via marriage/partnerships. Of course this design may confound the effect of differences in signaling sexual orientation and differences in the effects of orientation between single and married/registered women.

two pairs of applicants to employers in such a manner as to first provide information on the effect of masculinity among straight women, and then on sexual orientation. All women describe themselves as single.

Weichselbaumer (2003) finds that lesbian women received substantially fewer callbacks, and this was unrelated to masculinity or femininity.[77] She argues that statistical discrimination is not likely to underlie the results. First, as emphasized in Zschirnt and Ruedin (2016), the detailed applications in German-speaking labor markets leave less scope for this kind of discrimination. Second, because lesbians are raised in the same kinds of families and environments as heterosexuals, there is less reason to expect unobserved differences like those we might assume disadvantage minorities (although there could be nonrandom selection on which lesbians are "out"). Third, because of the lower likelihood of having children (especially in earlier years), if anything, we might expect statistical discrimination to favor lesbians over heterosexual women.

Drydakis (2009) used a similar strategy to study discrimination against gays in Greece, although without attention to gender stereotypes. Rather, to counter gay stereotypes regarding femininity, the gay and straight resumes included similar hobbies that, according to the author, implied similar masculinity. He finds strong evidence of hiring discrimination against gays, with callback rates lower by about 25 percentage points. In an interesting extension, Drydakis had fictitious applicants field return phone calls to try to get information on wages offered. The paper only reports wage regressions for pairs where both applicants received wage offers, and in these cases there was no detectable wage offer difference. In Drydakis's (2011) very similar study of discrimination against lesbian women in Greece, which finds much lower callback rates for lesbians, he reports all of the evidence on wage offers, finding a wage penalty for lesbians exceeding eight percent in the sample where only one or the other gets an offer, but only 4.8 percent in the matched sample where both get callbacks. These results suggest that the finding of equal wage offers for gay men in the first study may reflect the selection on pairs where both men received callbacks.

The wage evidence for women contrasts with the usual finding from nonexperimental data that lesbian women earn a wage premium. This could be due to differences in whether sexual orientation is known to the employer (as Weichselbaumer 2003 suggested), differences in wage offers measured by fielding phone calls versus actual accepted wage offers, or differences in wages based on gender stereotype (the resumes in this study did not particularly try to signal more masculine stereotypes).

Tilcsik (2011) presents the first large-scale correspondence study of discrimination against gay men in the United States, which includes some unique and potentially valuable features. First, it looks at variation in whether there is a law barring discrimination based on sexual orientation (looking at the seven states the study covers, as well as city and county laws). Second, Tilcsik examines the association between outcomes and state-level measures of attitudes toward gays. Third, the study uses text from ads to identify those requesting stereotypically male traits, to see if gays are less likely to get callbacks for these ads—presumably because of more feminine stereotypes, although Tilcsik does not manipulate the resumes to generate variation in masculinity among applicants.[78]

[77] Weichselbaumer (2004) uses these same data to study sex discrimination among heterosexuals, concluding that it is not driven by whether women have more masculine or feminine identities.

[78] To do this, ads have to be matched to specific callbacks from employers. In Neumark, Burn, and Button

The study finds a lower callback rate for gay applicants—7.2 percent, versus 11.5 percent for straight applicants. Tilcsik's (2011) multivariate analysis suggests that antidiscrimination laws may reduce discrimination, although the evidence does not uniformly point in this direction, and Tilcsik is asking a lot of the data with only seven states. There is also some evidence that public support for gay rights is associated with less discrimination, although it is hard to untangle this from the effect of laws that are passed. The evidence also suggests that gays were less likely to receive callbacks from employers with ads listing stereotypically male traits.

There are two other field experiments on discrimination based on sexual orientation in the United States. Bailey, Wallace, and Wright (2013) used triplets of resumes with a straight male, a straight female, and a gay or lesbian applicant (with sex alternating). And Mishel (2016) studied women only. Both studies used gay/lesbian/LBGT organizational leadership or membership to indicate sexual orientation. Bailey et al. found no discrimination against gay or lesbian applicants, whereas Mishel found significantly lower callback rates (by about one-third) for lesbian/queer women.

Ahmed, Andersson, and Hammarstedt (2013) conducted a correspondence study in Sweden of discrimination against both gays and lesbians. This study signals sexual orientation via volunteer work in an organization, and there was no manipulation of gender identity. The authors focus on many occupations, and, interestingly, find evidence of discrimination against gays only in male-dominated occupations, and against lesbians only in female-dominated occupations. This evidence may imply that discrimination based on sexual orientation depends more on gender stereotypes than on sexual orientation per se, although that may be more true in a country like Sweden, with less negative attitudes toward gays and lesbians than other countries (and indeed the overall evidence is weaker than in other countries).

Finally, Drydakis's (2014) correspondence study for Cyprus also focuses on both gays and lesbians. The methods and findings are similar to his studies for Greece, with regard to both hiring and potential wage offers. A unique feature of this study is its attempt to test for statistical versus taste discrimination via the provision, for some applicants, of additional information on grades, positive personality traits (e.g., efficient, organized), and reference letters attesting to work commitment, no absenteeism, etc. However, we should keep in mind Weichselbaumer's (2003) argument that there is no clear reason for statistical discrimination along these dimensions for gays and lesbians, which may explain the findings that discrimination against gays and lesbians was similar for the more- and less-informative applications.

The experimental literature on discrimination against gays and lesbians generally finds evidence of discrimination, although there may be less discrimination in more recent studies (see also Baert 2014). Whether this has to do with study design, the country studied, or increased acceptance of gays and lesbians is an interesting question that can only be answered by comparing similar studies over time and across locations, although there may not yet be enough studies to draw firm conclusions.

7.7.5 *Disabled*

There are a few correspondence studies of discrimination against the disabled. Ravaud, Madiot, and Ville (1992) sent unsolicited applications to French companies, differentiating applicants by a sentence in the cover letter indicating that an applicant was paraplegic (i.e., in a wheelchair) as a result of an accident

(forthcoming) we discuss the difficulties of doing these matches for phone responses. Tilcsik (2011) does not indicate how this problem was handled.

in 1982. Applicants were also distinguished by high or low qualifications based on education. The nondisabled applicants were much more likely to receive favorable responses—1.78 times more likely for the higher-qualified applicants, and 3.2 times more likely for the lower-qualified applicants. Ravaud et al. argue that their approach rules out the possibility that the disability they study affects productivity directly, although it is hard to be definitive about this. In particular, they argue that the disability on which they focus has no other interactions (by which I think they mean with other types of physical problems or disease), and that by dating the accident to the end of general education, training would have occurred after the disability occurred. Moreover, the applications indicated that the disabled applicants (and the nondisabled) had a driver's license, suggesting no concerns about this kind of mobility.

Baert (2016) studies disability discrimination in Belgium. The disabilities considered are blindness, deafness, and autism. The disabilities are noted in the cover letter and the resume, along with language about how the worker accommodates (e.g., a guide dog), and a statement that "my disability does not make me less productive." The disabled candidates were about half as likely to receive a positive callback.[79] As Baert notes, it is hard to know whether to interpret the difference as discrimination, given potential productivity differences, although he asserts that he assigned disability to occupations where disabled workers would be equally productive, perhaps after reasonable workplace adjustments (e.g., accountant for blind applicants, and carpenter for deaf applicants). The author seems to have information on the reasons employers give for positive or non-positive callbacks (which apparently is common, and also legally obligatory in Belgium), and employers rarely point to the disability. But it is not clear that they would honestly reveal this information. Finally, Baert also implements the correction for differences in the variances of unobservables, using the identifying assumption that distance from the worker's residence to the workplace (which is significantly negatively associated with callbacks) has an equal effect for the different groups, and reports that this correction does not alter the conclusions.

Finally, Ameri et al. (2018) conducted a larger-scale correspondence study in the United States. They focus on positions in accounting, and two different disabilities—spinal cord injuries, and Asperger's syndrome—which the authors argue would not affect productivity in this field. Information on these disabilities was revealed in the cover letter, with all cover letters indicating volunteer work for a disability organization, and the disabled applicants signaling their disability directly, along with a statement that "my disability does not interfere with my ability to perform the skills needed in a finance environment." The results point to significantly lower callbacks for disabled applicants (26 percent lower, on a base response rate to the non-disabled of 6.6 percent). The gap was larger for the more skilled applicants, and small and not statistically significant for the less experienced applicants. The type of disability made little difference.

Ameri et al. also provide evidence on the effects of antidiscrimination laws. Noting that both disabilities are clearly covered by the Americans with Disabilites Act (ADA), they break out results based on whether private employers have fewer than fifteen employees and hence are not covered by the ADA. Second, they explore whether employers were in states where state disability laws extend to smaller employers.[80] A variety of

[79] The study also focuses on the effect of a wage subsidy for hiring disabled workers.

[80] See Neumark, Song, and Button (2017) for non-experimental evidence on this policy variation and hiring of

analyses confirm the apparent effect of the ADA at the fifteen cutoff, but there is not consistent evidence of effects of state laws. It is also possible that large employers can more easily accommodate disabled workers.

7.7.6 *Long-term Unemployed*

After the Great Recession in the United States, unemployment durations increased, particularly for older workers (Neumark and Button 2014). This raised concerns of "discrimination" against the long-term unemployed, even leading to President Obama proposing, in the American Jobs Act, to prohibit employers from discriminating against unemployed workers when hiring.[81] These developments prompted the extension of AC studies to the difficulties of the long-term unemployed in finding new jobs, with particular attention paid to the question of what long-term unemployment signals to employers.

Kroft, Lange, and Notowidigdo (2013) conducted a correspondence study across 100 US cities. They found that applicants with long-term unemployment spells were much less likely to receive callbacks. (In a similar study, Ghayad (n.d.) similarly documents lower hiring of those in long-term unemployment spells.) Kroft et al. also find that while employers are less likely to call back those unemployed for a longer spell, the stigmatizing effect of a long unemployment spell is weaker in slacker labor markets, and stronger in tight labor markets, consistent with employers viewing long-term unemployment as a more negative signal when it is easier to find jobs.

Eriksson and Rooth (2014) also find that, in Sweden, fictitious applicants currently in long unemployment spells are less likely to get callbacks. They also study the longer-term stigmatizing effect of unemployment, and find no adverse effect on callbacks of long unemployment spells in the past, suggesting that any negative signals or loss of skills from earlier spells are offset by more recent experience. This is a particularly nice application of the correspondence study technique, because sorting out the effect of past unemployment spells on current labor market outcomes poses a severe challenge of controlling for heterogeneity, especially in models for the length of current unemployment spells.

To be clear, the authors of these studies interpret this evidence in the context of understanding duration dependence in unemployment spells, rather than discrimination (and as a result, I do not include the studies discussed in this subsection in table 3). Nonetheless, the application of the experimental methods from the discrimination literature is interesting. And the evidence in Kroft, Lange, and Notowidigdo (2013), in particular, fits to some extent into the broader issue of statistical discrimination, in that it seems to confirm that employers have imperfect information about applicants and, in this case, use external information to draw inferences about them. However, Farber, Silverman, and von Wachter (2017), in their correspondence study of women in the US labor market, do not find evidence of lower callback rates for those with long unemployment durations. (Kroft et al. find this effect more strongly for women than for men, while Ghayad (n.d.) uses male applicants only.) Farber et al. suggest that the most likely source of the difference is that Kroft et al. cover considerably younger ages, and for younger applicants where the work history is less informative, employers may place more weight (in tight labor markets) on the length of unemployment duration. Yet they also note that another recent correspondence study on unemployment durations for the United States, by Nunley et al.

the disabled.

[81] See https://www.whitehouse.gov/the-press-office/2011/09/08/fact-sheet-american-jobs-act (viewed August 23, 2016).

(2017), focuses on relatively recent college graduates and finds no effect of unemployment duration on callbacks.

7.8 *Wages and Other Outcomes*

AC studies of labor market discrimination focus almost exclusively on hiring, as the preceding extensive review demonstrates. However, much of the nonexperimental literature on discrimination is about wages, and in that literature the problem of group-related unobservables looms large. Can AC studies leverage their ability to control for unobservables and also tell us more about other labor market outcomes—in particular, whether wage offers or other benefits reflect discrimination? Audit studies can do this in principle, since testers actually receive job offers. Yet very few studies have tried to capture wage outcomes, and none have done so convincingly.

Bendick, Brown, and Wall (1999) report evidence of potentially worse job characteristics offered to older testers who received job offers, but based on very few observations. Pager, Bonikowski, and Western (2009) report some evidence of "channeling" of black and Latino applicants, during the interview, to lower-level jobs than those for which ads were posted. As noted above, Drydakis's (2009, 2011) correspondence studies on discrimination against gays and lesbians tried to get information on potential wage offers from callbacks, although it is not clear what these mean in the context of correspondence studies, in part because not all callbacks result in a job offer.

Finally, note that answering the question of wages offered in AC tests is different from the question raised by Heckman (1998) as to whether evidence of hiring discrimination necessarily implies that a discriminatory wage gap will result, which is a question about the effect of discrimination at the market level. It may be that the best answer to this market-level question, to date, comes from Charles and Guryan's (2008) work on discriminatory attitudes and wage differentials. Nonetheless, it would be useful to know just how strongly the evidence on hiring discrimination from the large body of AC studies helps to explain wage gaps that, in nonexperimental studies, are often viewed through the lens of discrimination.

7.9 *Anonymized Job Applications*

An interesting and different kind of experimental method to study discrimination in hiring is to anonymize applicants and see if selection for interviews (and potentially hiring) is affected. This parallels the Goldin and Rouse (2000) study of orchestras, albeit with experimentally induced variation. This anonymization would presumably only work for selection for interviews. While past research has indicated that most of the discrimination seems to occur at this stage, that could change if the ability to discriminate in selecting candidates for interviews was eliminated, although perhaps not if it is easier to enforce nondiscrimination in hiring from a pool of interviewees.

Krause, Rinne, and Zimmermann (2012) studied economics PhD applicants to a European research institute, which on its own randomized the anonymization of demographic information before the applications went to the hiring committee.[82] After the hiring process was complete, applicants were asked for permission to use their data in the study; 65 percent agreed. The key result is that female applicants were more likely to be invited for interviews in the non-anonymous sample, whereas this advantage was erased

[82] This kind of study echoes the call in Bendick and Nunes (2012) for more experimental research on strategies to reduce discrimination, which likely have to be undertaken by employers.

in the anonymous sample. Taken literally, this means that there was discrimination in favor of female candidates in the non-anonymous setting, which could not occur with anonymous applications. On the other hand, the hiring committee was aware of the experiment, and this could have motivated them to be nondiscriminatory in evaluating the non-anonymous applications.[83] There was, however, no such pattern regarding non-Western applications, although that may be because discrimination in this case is not as much of a social taboo (perhaps because of the demographic composition of the research staff at the institute).[84] Regardless, at this point this study could be viewed as providing additional confirmation of an absence of discrimination against women from experimental methods.

A second application of this method, to hiring of minority job candidates in France, found that minorities fared worse under anonymization, getting a smaller share of interviews (Behaghel, Crépon, and Le Barbanchon 2015). (The hiring gap also widened, but not as much as the interview gap.) This conclusion contrasts sharply with the results from AC studies, although the authors suggest reasons why this experiment may give misleading results. In this study, the French public employment service offered firms the choice of participating, in which case they were randomized to receive standard or anonymized job applications. Among the participating firms, those that received anonymous applications interviewed fewer minority candidates. The authors suggest that this happens because, under anonymization, negative characteristics of job applicants, which are more likely to be associated with minority group membership but that firms might downweight for minorities, have a larger effect than when minority status is revealed. They find evidence consistent with this interpretation, based on ratings of applicants in both treatments that they obtained from counselors from the public employment service.[85]

Behaghel, Crépon, and Le Barbanchon (2015) note that the firms that agreed to participate were similar on most observables, except that they hired more minorities. When they joined the study, then, those that received anonymous applications may have been unable to continue preferential hiring of minorities. Under this interpretation, extending anonymization to all employers would have ambiguous effects, depending on to what extent the nonparticipating employers engage in discrimination against minorities, and to what extent anonymization prevents this.[86] The considerations raised here suggest that while anonymization of resumes is potentially promising as a research strategy, it will need to avoid experimenter effects. At this point, the far greater body of research on discrimination using AC studies provides more convincing evidence. By the same token, from a policy perspective, I would caution against concluding, yet, that anonymization of resumes will not

[83] Behaghel, Crépon, and Le Barbanchon (2015), discussed below, refer to this as the "John Henry" effect.

[84] Åslund and Nordström Skans (2012) present evidence on anonymized job application procedures from nonexperimental data in Sweden, in which they find increased interviews and job offers for women, but only increased interviews for ethnic minorities.

[85] An alternative explanation is the John Henry effect mentioned above, whereby the controls change their behavior to avoid appearing discriminatory, which the treatments cannot do because their job applications are anonymous. The authors suggest they can rule this out based on no difference in behavior among the controls during the experiment and after, when they are no longer being observed. It would be better to have data from before and during the experiment, because the experiment could have a residual effect.

[86] This same issue arose in the Agan and Starr (2018) study discussed above, where restricting information on criminal background seems to have increased statistical discrimination against blacks.

reduce discrimination; the results could well differ in other contexts.

8. *Conclusions and Directions for Future Research*

8.1 *What Have We Learned from this Vast Body of Experimental Research on Labor Market Discrimination?*

We have learned that for most groups for which nonexperimental data on wages is consistent with labor market discrimination, much of the experimental research provides confirming evidence of hiring discrimination. This is most evident with respect to race and ethnicity, perhaps because there is so much more evidence.[87] There is also pervasive evidence consistent with discrimination against gays, criminals, and the disabled. In addition, the experimental research points to hiring discrimination against groups for which the prima facie evidence (most notably, for wages) does not point to discrimination—in particular, lesbians and older workers (at least older women).

The audit and correspondence study evidence on sex discrimination provides an interesting exception. Despite a great deal of nonexperimental evidence consistent with sex discrimination, the experimental research does not give a clear indication that women are treated less favorably in hiring decisions. Indeed, the evidence appears to point more to discrimination against women in some jobs and men in other jobs, in a manner that helps replicate the existing sex segregation of jobs. If this is in fact what happens, it poses a bit of a puzzle for labor economists, since our models of discrimination do not naturally predict this pattern. Perspectives on discrimination from other fields, such as those emphasizing norms regarding who does which job, might fit these facts better. And the experimental evidence does not necessarily imply that the unexplained wage gaps between men and women do not reflect discrimination.

8.2 *What Are the Policy Implications?*

Overall, though, this review reinforces the conclusion that hiring discrimination is pervasive. This may be, in part, because even with aggressive antidiscrimination laws such as in the United States, it is hard to root out discrimination in hiring. Enforcement of antidiscrimination laws in the United States relies on the legal process, and hence on potential rewards to plaintiffs' attorneys. In hiring cases it is difficult to identify a class of affected workers, inhibiting class-action suits and thus substantially limiting awards. In addition, economic damages can be small in hiring cases because one employer's action may extend a worker's spell of unemployment only modestly. (Terminations, in contrast, can entail substantial lost earnings and benefits, and, for older workers, significant lost pension accruals.)

What does this evidence, and these considerations, imply for policy? First, if discrimination along the hiring dimension is particularly hard to root out, the evidence of hiring discrimination may not carry over to other labor market decisions (e.g., pay, promotions, and layoffs). Thus, the evidence in this review should be taken primarily as posing the challenge of how to reduce hiring discrimination.

More substantively with respect to policy, the pervasive evidence of discrimination in hiring, coupled with reasons to believe the current enforcement mechanisms may be inadequate, may have a few important policy implications. First, we may need to consider policies to reduce hiring discrimination that

[87] Recall, however, the finding that evidence of race and ethnic labor market discrimination sometimes weakens substantially when the estimates are corrected for bias from different variances of the unobservable (Neumark and Rich 2016). Additional approaches to the Heckman critique may therefore prove instructive in sharpening our interpretation of this evidence.

are less reliant on attorneys and plaintiffs, and perhaps more proactive, rather than reactive. Affirmative action is of course such a policy, but in the United States its scope is limited to federal contractors.

Second, if it is true that hiring discrimination is particularly impervious to enforcement via private lawsuits, the EEOC and state Fair Employment Practices agencies may have to do more to focus their resources on discrimination in hiring. On this score, it is noteworthy that the EEOC's recent strategic enforcement plan (for 2013–16) lists, as its first priority, targeting "class-based recruitment and hiring practices that discriminate against racial, ethnic and religious groups, older workers, women, and people with disabilities."[88] This has been interpreted as a change in EEOC policy that will increase the focus on hiring patterns and reduce the focus on individual claims of discrimination[89]—exactly what may be needed to better root out discrimination in hiring.

Finally, the appropriate policy response to statistical discrimination is likely quite different from the response to taste discrimination. Although current law makes both types of discrimination illegal, the law is very much focused on the cost approach that seems to directly target taste discrimination, as opposed to perhaps considering other policies mandating how screening and evaluation of candidates is done that might reduce statistical discrimination with reduced reliance on ex post enforcement and the courts. An example is recent "ban-the-box" policies that eliminate the collection of information on criminal background at the earliest screening stage, which can be interpreted as encouraging employers to rely less on this cheap screen at the outset and instead to evaluate the candidate—including his or her criminal record—at a later stage when the employer also has other information. Affirmative action policies may also act in this fashion, reducing reliance on cheap screens like race (Holzer and Neumark 2000, Miller 2017).[90] At the same time, the anonymization studies caution that removing group identifiers from applications—which may seem like an intuitively appealing way to eliminate discrimination—can cut against these efforts. And if we add implicit discrimination to the mix, the best policy response may be different still.

8.3 *What Are the Important Questions for Future Research?*

There are open questions about precisely what we learn from the experimental evidence on some groups, which probably revolve mainly around the question of "is the differential treatment discrimination?" This arises most sharply, I think, with respect to research on those with a criminal background, and the disabled, who—for very different reasons—might be viewed as having different productivity. With regard to the disabled, further progress on this question likely requires incorporating more information on the relationships between specific disabilities and productivity.

Many papers using experimental methods to study labor market discrimination have tried to extend their results to uncover the nature of discrimination. This is a critical question, as it can inform how both policy makers and employers might respond to

[88] See https://www.eeoc.gov/eeoc/plan/sep.cfm (viewed August 18, 2016).

[89] See http://www.lexology.com/library/detail.aspx?g=b2bac38f-9380-4b24-9c5a-b16e921daba2 (viewed August 18, 2016).

[90] Both of these studies present evidence that affirmative action is associated with firms investing more in screening, training, and evaluation of employees, including once they begin work. Miller argues that this helps explain his finding that the relative rate of black hiring at firms bound by affirmative action continued to increase after they became bound (because of federal contractor status).

reduce discrimination. More work is needed to pin down compelling tests, and to replicate these results across studies. In particular, I have explained why I find many studies trying to distinguish between statistical and taste discrimination less than convincing, because we do not know—to paraphrase Senator Howard Baker's famous question at the Watergate hearings—"what does the employer know, and when does he know it?"[91] But that should be taken as a challenge to come up with more compelling tests, perhaps by supplementing field experiments with survey or interview evidence on what employers know, what information they use, what assumptions they make, etc., and how this changes during the recruitment, interview, and hiring process. This strategy parallels recent work such as Rooth (2010) in incorporating supplemental information on participants in field experiments. This is not easy, and is likely to require studies of smaller numbers of job applicants with more intensive efforts to enlist employers in the research endeavor.

Moreover, the evidence on the nature of discrimination that does exist is all over the map, with no clear indication from the existing studies. There is some evidence that supports each of the three central explanations of discrimination (taste, statistical, and implicit), some evidence against both statistical and taste discrimination, and a number of places where conclusions drawn in support of one of these explanations may not be well founded. This is likely a reflection of two things. First, all three types of discrimination may matter, and perhaps differently for different groups. There may, for example, be more taste discrimination against groups whose "otherness" is most pronounced (say, Middle Eastern immigrants), and more statistical discrimination against groups with well-known but hard to predict differences (e.g., women of childbearing age). Second, distinguishing among the competing explanations is simply difficult.

Researchers have frequently noted that audit and correspondence studies have somewhat limited scope—applying only to certain segments of the labor market, excluding some common channels of job search, etc. But there has been virtually no work trying to understand what these limitations might mean, or how we might expand the boundaries of experimental studies to reduce these limitations. Can we do more to figure out the role of the kinds of online job applications now used in field experiments in the overall job search of workers, perhaps in part by exploiting cross-country differences in reliance on government employment offices versus more private search? With changes in the technology of how people search for jobs, can we expand the types of jobs covered? Can social media be harnessed to improve these studies? (Alternatively, does the ability to use social media and other information technology to check credentials and identities undermine future field experiments?[92])

[91] There is some quite compelling evidence of statistical discrimination in other markets (e.g., Castillo and Petrie 2010, Laouenan and Rathelot 2015, List 2004), made easier in some cases because of the availability of information on learning about market participants (as in Laouenan and Rathelot's study).

[92] If, for example, a correspondence study uses minority applicants with high credentials relative to the population, employers may be more likely to try to verify these credentials for minorities, and failing to do so may be less likely to call them back, leading to an overestimate of discrimination. In their study of race, ethnicity, and criminal background, Pager, Bonikowski, and Western (2009) set up voicemail boxes to see if employers checked references. Almost none did (four employers out of 350 that were tested). This finding may not generalize, perhaps especially with respect to higher-skilled jobs when the stakes may be higher for employers, and more information available. That said, these methods have rarely been used to study high-skilled jobs. Nonetheless, the evidence available thus far indicates that at least for the kinds of studies generally conducted, concerns that AC studies might be invalidated because of employers having more access to information to check on the resumes may be largely unfounded.

And what about the other side of the market? How important is the kind of hiring we study, relative to the hiring employers do, and does the behavior these experiments study accurately capture a large extent of the hiring that at least these same employers—not to mention other employers—do? In general, researchers doing field experiments on labor market discrimination might consider how to assess the representativeness of the results from these experiments, and try to extend such experiments into a more diverse set of jobs and job search methods.

The focus on the application and hiring stage in field experiments on discrimination typically precludes gathering evidence on other types of discriminatory behavior—perhaps most importantly pay. There have been a few recent efforts to try to incorporate information on pay-related outcomes into field experiments, but nothing like proceeding all the way to final wage offers in an audit study. A recent experimental study of race discrimination in consumer markets goes all the way through to transaction prices (Doleac and Stein 2013). However, that study could consummate the actual transactions, so there presumably was not a problem of getting human subjects' approval. It is an open question whether an institutional review board would currently approve even a standard audit study (as opposed to a correspondence study), let alone one that extended the interview process to try to settle on pay.

Finally, I find the research on policy interventions to reduce discrimination intriguing, even if the two field experiments discussed (Krause, Rinne, and Zimmermann 2012; Behaghel, Crépon, and Le Barbanchon 2015) are so far discouraging in suggesting that women or minorities did worse following these interventions. The contexts studied—a research institute that may have been engaging in affirmative action, and an employment service into which there may have been self-selection of firms—may have been unique, and more evidence from run-of-the-mill employers would clearly be of value. Alternatively, the research strategy of using field experiments before and after broader policy changes may prove informative, although requiring some serendipity in the timing of such studies.

Based on these considerations, if I were advising a highly ambitious graduate student who wanted to do a significant audit or correspondence study, I would discourage them from simply doing another study of differential treatment in hiring between two groups—even if they were groups that have not been studied much. Such a study would, to a large extent, provide another data point in a field where I think we know where most data points are going to lie—although such work can play a role in advocacy or even litigation on behalf of such groups. Rather, I think the important gains are going to come from embedding an AC study in a research project that is going to shed additional light that helps us understand the findings—whether concerning the nature of discrimination, the implications of the experiment for the broader labor market, or perhaps how policy variation or other differences impact the discrimination these experiments are intended to measure.

References

Agan, Amanda, and Sonja Starr. 2018. "Ban the Box, Criminal Records, and Statistical Discrimination: A Field Experiment." *Quarterly Journal of Economics* 133 (1): 191–235..

Agerström, Jens, and Dan-Olof Rooth. 2011. "The Role of Automatic Obesity in Real Hiring Discrimination." *Journal of Applied Psychology* 96 (4): 790–805.

Ahmed, Ali M., Lina Andersson, and Mats Hammarstedt. 2013. "Are Gay Men and Lesbians Discriminated against in the Hiring Process?" *Southern Economic Journal* 79 (3): 565–85.

Aigner, Dennis J., and Glen G. Cain. 1977. "Statistical Theories of Discrimination in Labor Markets." *ILR Review* 30 (2): 175–87.

Altonji, Joseph G., and Charles R. Pierret. 2001. "Employer Learning and Statistical Discrimination." *Quarterly Journal of Economics* 116 (1): 313–50.

Ameri, Mason, Lisa Schur, Meera Adya, Scott Bentley, Patrick McKay, and Douglas Kruse. 2018. "The Disability Employment Puzzle: A Field Experiment on Employer Hiring Behavior." *Industrial and Labor Relations Review* 71 (2): 329–64.

Anderson, Lisa R., Roland G. Fryer, Jr., and Charles A. Holt. 2006. "Discrimination: Experimental Evidence from Psychology and Economics." In *Handbook on the Economics of Discrimination*, edited by William M. Rogers, 97–115. Cheltenham and Northampton: Edward Elgar.

Andreoni, James, and Ragan Petrie. 2008. "Beauty, Gender and Stereotypes: Evidence from Laboratory Experiments." *Journal of Economic Psychology* 29 (1): 73–93.

Arrow, Kenneth J. 1972. "Some Mathematical Models of Racial Discrimination in the Labor Market." In *Racial Discrimination in Economic Life*, edited by Anthony H. Pascal, 187–204. Lexington: Lexington Books.

Arrow, Kenneth J. 1998. "What Has Economics to Say About Racial Discrimination?" *Journal of Economic Perspectives* 12 (2): 91–100.

Åslund, Olof, and Oskar Nordström Skans. 2012. "Do Anonymous Job Application Procedures Level the Playing Field?" *ILR Review* 65 (1): 82–107.

Ayres, Ian, and Peter Siegelman. 1995. "Race and Gender Discrimination in Bargaining for a New Car." *American Economic Review* 85 (3): 304–21.

Baert, Stijn. 2014. "Career Lesbians: Getting Hired for Not Having Kids?" *Industrial Relations Journal* 45 (6): 543–61.

Baert, Stijn. 2015. "Field Experimental Evidence on Gender Discrimination in Hiring: Biased as Heckman and Siegelman Predicted?" *Economics: The Open Access, Open-Assessment E-Journal*. https://biblio.ugent.be/publication/6912176/file/6912177.pdf.

Baert, Stijn. 2016. "Wage Subsidies and Hiring Chances for the Disabled: Some Causal Evidence." *European Journal of Health Economics* 17 (1): 71–86.

Baert, Stijn, Bart Cockx, Niels Gheyle, and Cora Vandamme. 2015. "Is There Less Discrimination in Occupations Where Recruitment Is Difficult?" *ILR Review* 68 (3): 467–500.

Baert, Stijn, and Ann-Sophie De Pauw. 2014. "Is Ethnic Discrimination Due to Distaste or Statistics?" *Economics Letters* 125 (2): 270–73.

Baert, Stijn, Ann-Sophie De Pauw, and Nick Deschacht. 2016. "Do Employer Preferences Contribute to Sticky Floors?" *ILR Review* 69 (3): 714–36.

Baert, Stijn, and Elsy Verhofstadt. 2015. "Labour Market Discrimination against Former Juvenile Delinquents: Evidence from a Field Experiment." *Applied Economics* 47 (11): 1061–72.

Bailey, John, Michael Wallace, and Bradley Wright. 2013. "Are Gay Men and Lesbians Discriminated against When Applying for Jobs? A Four-City, Interest-Based Field Experiment." *Journal of Homosexuality* 60 (6): 873–94.

Banerjee, Abhijit, Marianne Bertrand, Saugato Datta, and Sendhil Mullainathan. 2009. "Labor Market Discrimination in Delhi: Evidence from a Field Experiment." *Journal of Comparative Economics* 37 (1): 14–27.

Bartolucci, Cristian. 2013. "Gender Wage Gaps Reconsidered: A Structural Approach Using Matched Employer–Employee Data." *Journal of Human Resources* 48 (4): 998–1034.

Bartolucci, Cristian. 2014. "Understanding the Native–Immigrant Wage Gap Using Matched Employer–Employee Data: Evidence from Germany." *ILR Review* 67 (4): 1166–202.

Bartoš, Vojtěch, Michal Bauer, Julie Chytilová, and Filip Matějka. 2016. "Attention Discrimination: Theory and Field Experiments with Monitoring Information Acquisition." *American Economic Review* 106 (6): 1437–75

Baum, Charles L., II, and William F. Ford. 2004. "The Wage Effects of Obesity: A Longitudinal Study." *Health Economics* 13 (9): 885–99.

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

Behaghel, Luc, Bruno Crépon, and Thomas Le Barbanchon. 2015. "Unintended Effects of Anonymous Résumés." *American Economic Journal: Applied Economics* 7 (3): 1–27.

Bendick, Marc, Jr., Lauren E. Brown, and Kennington Wall. 1999. "No Foot in the Door: An Experimental Study of Employment Discrimination against Older Workers." *Journal of Aging and Social Policy* 10 (4): 5–23.

Bendick, Marc, Jr., Charles W. Jackson, and Horacio Romero. 1997. "Employment Discrimination against Older Workers." *Journal of Aging and Social Policy* 8 (4): 25–46.

Bendick, Marc, Jr., and Ana P. Nunes. 2012. "Developing the Research Basis for Controlling Bias in Hiring." *Journal of Social Issues* 68 (2): 238–62.

Berman, Phyllis W., Barbara A. O'Nan, and Wayne Floyd. 1981. "The Double Standard of Aging and the Social Situation: Judgments of Attractiveness of the Middle-Aged Woman." *Sex Roles* 7 (2): 87–96.

Bertrand, Marianne, Dolly Chugh, and Sendhil Mullainathan. 2005. "Implicit Discrimination." *American Economic Review* 95 (2): 94–98.

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.

Biddle, Jeff E., and Daniel S. Hamermesh. 2013. "Wage Discrimination over the Business Cycle." *IZA Journal of Labor Policy* 2 (7).

Black, Dan A. 1995. "Discrimination in an Equilibrium Search Model." *Journal of Labor Economics* 13 (2): 309–34.

Blau, Francine D., and Lawrence M. Kahn. 2017. "The Gender Wage Gap: Extent, Trends, and Explanations." *Journal of Economic Literature* 55 (3): 789–865.

Blommaert, Lieselotte, Marcel Coenders, and Frank van Tubergen. 2014. "Ethnic Discrimination in Recruitment and Decision Makers' Features: Evidence from Laboratory Experiment and Survey Data using a Student Sample." *Social Indicators Research* 116 (3): 731–54.

Booth, Alison L., Andrew Leigh, and Elena Varganova. 2012. "Does Ethnic Discrimination Vary across Minority Groups? Evidence from a Field Experiment." *Oxford Bulletin of Economics and Statistics* 74 (4): 547–73.

Burn, Ian. 2015. "Legal Differences in State Non-discrimination Laws and the Effect of Employment Protections for Gay Men." Unpublished.

Büsch, Victoria, Svenn-Åge Dahl, and Dennis A. V. Dittrich. 2009. "An Empirical Study of Age Discrimination in Norway and Germany." *Applied Economics* 41 (5): 633–51.

Caliendo, Marco, and Markus Gehrsitz. 2015. "Obesity and the Labor Market: A Fresh Look at the Weight Penalty." *Economics & Human Biology* 23: 209–25.

Carlsson, Magnus. 2011. "Does Hiring Discrimination Cause Gender Segregation in the Swedish Labor Market?" *Feminist Economics* 17 (3): 71–102.

Carlsson, Magnus, Luca Fumarco, and Dan-Olof Rooth. 2013. "Artifactual Evidence of Discrimination in Correspondence Studies? A Replication of the Neumark Method." Institute for the Study of Labor Discussion Paper 7619.

Carlsson, Magnus, and Dan-Olof Rooth. 2012. "Revealing Taste-Based Discrimination in Hiring: A Correspondence Testing Experiment with Geographic Variation." *Applied Economics Letters* 19 (18): 1861–64.

Castillo, Marco, and Ragan Petrie. 2010. "Discrimination in the Lab: Does Information Trump Appearance?" *Games and Economic Behavior* 68 (1): 50–59.

Castillo, Marco, Ragan Petrie, Maximo Torero, and Lisa Vesterlund. 2013. "Gender Differences in Bargaining Outcomes: A Field Experiment on Discrimination." *Journal of Public Economics* 99: 35–48.

Charles, Kerwin Kofi, and Jonathan Guryan. 2008. "Prejudice and Wages: An Empirical Assessment of Becker's *The Economics of Discrimination*." *Journal of Political Economy* 116 (5): 773–809.

Coate, Stephen, and Glenn C. Loury. 1993. "Will Affirmative-Action Policies Eliminate Negative Stereotypes?" *American Economic Review* 83 (5): 1220–40.

Correll, Shelley J., Stephen Benard, and In Paik. 2007. "Getting a Job: Is There a Motherhood Penalty?" *American Journal of Sociology* 112 (5): 1297–338.

Cross, Harry, Genevieve M. Kenney, Jane Mell, and Wendy Zimmermann. 1990. *Employer Hiring Practices: Differential Treatment of Hispanic and Anglo Job Seekers*. Washington, DC: Urban Institute Press.

Daniel, William Wentworth. 1968. *Racial Discrimination in England: Based on the PEP Report*. London: Penguin.

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

Davison, Heather K., and Michael J. Burke. 2000. "Sex Discrimination in Simulated Employment Contexts: A Meta-analytic Investigation." *Journal of Vocational Behavior* 56 (2): 225–48.

Devine, Patricia G. 1989. "Stereotypes and Prejudice: Their Automatic and Controlled Components." *Journal of Personality and Social Psychology* 56 (1): 5–18.

Dickinson, David L., and Ronald L. Oaxaca. 2009. "Statistical Discrimination in Labor Markets: An Experimental Analysis." *Southern Economic Journal* 76 (1): 16–31.

Dittrich, Marcus, Andreas Knabe, and Kristina Leipold. 2014. "Gender Differences in Experimental Wage Negotiations." *Economic Inquiry* 52 (2): 862–73.

Doleac, Jennifer L., and Luke C. D. Stein. 2013. "The Visible Hand: Race and Online Market Outcomes." *Economic Journal* 123 (572): F469–92.

Dovidio, John F., Samuel L. Gaertner, Kerry Kawakami, and Gordon Hodson. 2002. "Why Can't We Just Get Along? Interpersonal Biases and Interracial Distrust." *Cultural Diversity and Ethnic Minority Psychology* 8 (2): 88–102.

Drydakis, Nick. 2009. "Sexual Orientation Discrimination in the Labour Market." *Labour Economics* 16 (4): 364–72.

Drydakis, Nick. 2011. "Women's Sexual Orientation and Labor Market Outcomes in Greece." *Feminist Economics* 17 (1): 89–117.

Drydakis, Nick. 2014. "Sexual Orientation Discrimination in the Cypriot Labour Market. Distastes or Uncertainty?" *International Journal of Manpower* 35 (5): 720–44.

Duguet, Emmanuel, and Pascale Petit. 2005. "Hiring Discrimination in the French Financial Sector: An Econometric Analysis on Field Experiment Data." *Annales d'Économie et de Statistique* 78: 79–102.

Eckel, Catherine. 2008. "The Gender Gap: Using the Lab as a Window on the Market." In *Frontiers in the Economics of Gender*, edited by Francesca Bettio and Alina Verashchagina, 223–42. New York and London: Taylor and Francis, Routledge.

Eriksson, Stefan, and Dan-Olof Rooth. 2014. "Do Employers Use Unemployment as a Sorting Criterion When Hiring? Evidence from a Field Experiment." *American Economic Review* 104 (3): 1014–39.

Exley, Christine L., Muriel Niederle, and Lise Vesterlund. 2016. "Knowing When to Ask: The Cost of Leaning-In." Harvard Business School Working Paper 16-115.

Farber, Henry S., Dan Silverman, and Till von Wachter. 2017. "Factors Determining Callbacks to Job Applications By the Unemployed: An Audit Study." *RSF: The Russell Sage Foundation Journal of the Social Sciences* 3 (3): 168–201.

Fershtman, Chaim, and Uri Gneezy. 2001. "Discrimination in a Segmented Society: An Experimental Approach." *Quarterly Journal of Economics* 116 (1): 351–77.

Figinski, Theodore F. 2017. "The Effect of Potential Activations on the Employment of Military

Reservists: Evidence from a Field Experiment." *ILR Review* 70 (4): 1037–56.
Finlay, Keith. 2007. "Effect of Employer Access to Criminal History Data on the Labor Market Outcomes of Ex-offenders and Non-offenders." In *Studies of Labor Market Intermediation*, edited by David H. Autor, 89–125. Chicago and London: University of Chicago Press.
Fix, Michael E., and Raymond J. Struyk, eds. 1993. *Clear and Convincing Evidence: Measurement of Discrimination in America*. Washington, DC: Urban Institute Press.
Flory, Jeffrey A., Andreas Leibbrandt, and John A. List. 2015. "Do Competitive Workplaces Deter Female Workers? A Large-Scale Natural Field Experiment on Job Entry Decisions." *Review of Economic Studies* 82 (1): 122–55.
Foster, Andrew D., and Mark R. Rosenzweig. 1993. "Information, Learning, and Wage Rates in Low-Income Rural Areas." *Journal of Human Resources* 28 (4): 759–90.
Galarza, Francisco B., and Gustavo Yamada. 2014. "Labor Market Discrimination in Lima, Peru: Evidence from a Field Experiment." *World Development* 58: 83–94.
Gallo, Edoardo, Thomas Grund, and J. James Reade. 2013. "Punishing the Foreigner: Implicit Discrimination in the Premier League Based on Oppositional Identity." *Oxford Bulletin of Economics and Statistics* 75 (1): 136–56.
Ghayad, Rand. n.d. "Escaping the Unemployment Trap: Does Industry-Specific Human Capital Matter? Evidence from a Field Experiment." Unpublished.
Glover, Dylan, Amanda Pallais, and William Pariente. 2017. "Discrimination as a Self-Fulfilling Prophecy: Evidence from French Grocery Stores." *Quarterly Journal of Economics* 123 (3): 1219–60.
Gneezy, Uri, John List, and Michal K. Price. 2012. "Toward an Understanding of Why People Discriminate: Evidence from a Series of Natural Field Experiments." National Bureau of Economic Research Working Paper 17855.
Gneezy, Uri, Muriel Niederle, and Aldo Rustichini. 2003. "Performance in Competitive Environments: Gender Differences." *Quarterly Journal of Economics* 118 (3): 1049–74.
Goldberg, Matthew S. 1982. "Discrimination, Nepotism, and Long-Run Wage Differentials." *Quarterly Journal of Economics* 97 (2): 307–19.
Goldin, Claudia, and Cecilia Rouse. 2000. "Orchestrating Impartiality: The Impact of 'Blind' Auditions on Female Musicians." *American Economic Review* 90 (4): 715–41.
Greenwald, Anthony G., and Mahzarin R. Banaji. 1995. "Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes." *Psychological Review* 102 (1): 4–27.
Gronau, Reuben. 1988. "Sex-Related Wage Differentials and Women's Interrupted Labor Careers—The Chicken or the Egg." *Journal of Labor Economics* 6 (3): 277–301.
Groshen, Erica L. 1991. "The Structure of the Female/Male Wage Differential: Is It Who You Are, What You Do, or Where You Work?" *Journal of Human Resources* 26 (3): 457–72.
Hamermesh, Daniel S. 2011. *Beauty Pays: Why Attractive People Are More Successful*. Princeton and Oxford: Princeton University Press.
Harrison, Glenn W., and John A. List. 2004. "Field Experiments." *Journal of Economic Literature* 42 (4): 1009–55.
Heckman, James J. 1998. "Detecting Discrimination." *Journal of Economic Perspectives* 12 (2): 101–16.
Heckman, James J., and Peter Siegelman. 1993. "The Urban Institute Audit Studies: Their Methods and Findings." In *Clear and Convincing Evidence: Measurement of Discrimination in America*, edited by Michael E. Fix and Raymond J. Struyk, 187–258. Washington, DC: Urban Institute Press.
Hedegaard, Morten, and Jean-Robert Tyran. 2018. "The Price of Prejudice." *American Economic Journal: Applied Economics* 10 (1): 40–63.
Heilman, Madeline E. 1984. "Information as a Deterrent against Sex Discrimination: The Effects of Applicant Sex and Information Type on Preliminary Employment Decisions." *Organizational Behavior and Human Performance* 33 (2): 174–86.
Hellerstein, Judith K., and David Neumark. 1999. "Sex, Wages, and Productivity: An Empirical Analysis of Israeli Firm-Level Data." *International Economic Review* 40 (1): 95–123.
Hellerstein, Judith K., and David Neumark. 2008. "Workplace Segregation in the United States: Race, Ethnicity, and Skill." *Review of Economics and Statistics* 90 (3): 459–77.
Hellerstein, Judith K., David Neumark, and Kenneth R. Troske. 1999. "Wages, Productivity, and Worker Characteristics: Evidence from Plant-Level Production Functions and Wage Equations." *Journal of Labor Economics* 17 (3): 409–46.
Helleseter, Miguel Delgado, Peter Kuhn, and Kailing Shen. 2014. "Employers' Age and Gender Preferences: Direct Evidence from Four Job Boards." Unpublished.
Holzer, Harry J. 1998. "Why Do Small Establishments Hire Fewer Blacks than Large Ones?" *Journal of Human Resources* 33 (4): 896–914.
Holzer, Harry J., and David Neumark. 2000. "What Does Affirmative Action Do?" *ILR Review* 53 (2): 240–71.
Hosoda, Megumi, Eugene F. Stone-Romero, and Gwen Coats. 2003. "The Effects of Physical Attractiveness on Job-Related Outcomes: A Meta-analysis of Experimental Studies." *Personnel Psychology* 56 (2): 431–62.
Hutchens, Robert. 1986. "Delayed Payment Contracts and a Firm's Propensity to Hire Older Workers." *Journal of Labor Economics* 4 (4): 439–57.
Ioannides, Yannis M., and Linda Datcher Loury. 2004. "Job Information Networks, Neighborhood Effects, and Inequality." *Journal of Economic Literature* 42 (4): 1056–93.

Issacharoff, Samuel, and Erica Worth Harris. 1997. "Is Age Discrimination Really Age Discrimination?: The ADEA's Unnatural Solution." *NYU Law Review* 72 (4): 780–840.

Jackson, Linda A. 1992. *Physical Appearance and Gender: Sociobiological and Sociocultural Perspectives*. Albany: State University of New York Press.

Johnson, William R., and Derek Neal. 1998. "Basic Skills and the Black–White Earnings Gap." In *The Black–White Test Score Gap*, edited by Christopher Jencks and Meredith Phillips, 480–97. Washington, DC: Brookings Institution Press.

Jowell, Roger, and Patricia Prescott-Clarke. 1970. "Racial Discrimination and White-Collar Workers in Britain." *Race & Class* 11 (4): 397–417.

Kaas, Leo, and Christian Manger. 2012. "Ethnic Discrimination in Germany's Labour Market: A Field Experiment." *German Economic Review* 13 (1): 1–20.

Kahn, Lawrence M. 2000. "The Sports Business as a Labor Market Laboratory." *Journal of Economic Perspectives* 14 (3): 75–94.

Kite, Mary E., Gary D. Stockdale, Bernard E. Whitley, Jr., and Blair T. Johnson. 2005. "Attitudes toward Younger and Older Adults: An Updated Meta-analytic Review." *Journal of Social Issues* 61 (2): 241–66.

Kochhar, Rakesh. 2008. "Latino Labor Report, 2008: Construction Reverses Job Growth for Latinos." Washington, DC: Pew Research Center.

Krause, Annabelle, Ulf Rinne, and Klaus F. Zimmermann. 2012. "Anonymous Job Applications of Fresh Ph.D. Economists." *Economics Letters* 117 (2): 441–44.

Krings, Franciska, Sabine Sczesny, and Annette Kluge. 2011. "Stereotypical Inferences as Mediators of Age Discrimination: The Role of Competence and Warmth." *British Journal of Management* 22 (2): 187–201.

Kroft, Kory, Fabian Lange, and Matthew J. Notowidigdo. 2013. "Duration Dependence and Labor Market Conditions: Evidence from a Field Experiment." *Quarterly Journal of Economics* 128 (3): 1123–67.

Kuhn, Peter, and Kailing Shen. 2013. "Gender Discrimination in Job Ads: Evidence from China." *Quarterly Journal of Economics* 128 (1): 287–336.

Lahey, Joanna N. 2008. "Age, Women, and Hiring: An Experimental Study." *Journal of Human Resources* 43 (1): 30–56.

Lang, Kevin, and Jee-Yeon K. Lehmann. 2012. "Racial Discrimination in the Labor Market: Theory and Empirics." *Journal of Economic Literature* 50 (4): 959–1006.

Lang, Kevin, Michael Manove, and William T. Dickens. 2005. "Racial Discrimination in Labor Markets with Posted Wage Offers." *American Economic Review* 95 (4): 1327–40.

Laouenan, Morgane, and Roland Rathelot. 2015. "Ethnic Discrimination in an Online Marketplace of Vacation Rentals." Unpublished.

Lazear, Edward P. 1979. "Why Is There Mandatory Retirement?" *Journal of Political Economy* 87 (6): 1261–84.

List, John A. 2004. "The Nature and Extent of Discrimination in the Marketplace: Evidence from the Field." *Quarterly Journal of Economics* 119 (1): 49–89.

López Bóo, Florencia, Martin A. Rossi, and Sergio S. Urzúa. 2013. "The Labor Market Return to an Attractive Face: Evidence from a Field Experiment." *Economics Letters* 118 (1): 170–72.

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

Lundborg, Petter, Paul Nystedt, and Dan-Olof Rooth. 2014. "Body Size, Skills, and Income: Evidence from 150,000 Teenage Siblings." *Demography* 51 (5): 1573–96.

Miller, Conrad. 2017. "The Persistent Effect of Temporary Affirmative Action." *American Economic Journal: Applied Economics* 9 (3): 152–90.

Mincer, Jacob. 1974. *Schooling, Experience, and Earnings*. New York: National Bureau of Economic Research.

Mishel, Emma. 2016. "Discrimination against Queer Women in the U.S. Workforce: A Résumé Audit Study." *Socius* 2.

Mobius, Markus M., and Tanya S. Rosenblat. 2006. "Why Beauty Matters." *American Economic Review* 96 (1): 222–35.

Moreno, Martin, Hugo Ñopo, Jaime Saavedra, and Maximo Torero. 2012. "Detecting Gender and Racial Discrimination in Hiring through Monitoring Intermediation Services: The Case of Selected Occupations in Metropolitan Lima, Peru." *World Development* 40 (2): 315–28.

Mujcic, Redzo, and Paul Frijters. 2013. "Still Not Allowed on the Bus: It Matters If You're Black or White!" Institute for the Study of Labor Discussion Paper 7300.

Neal, Derek A., and William R. Johnson. 1996. "The Role of Premarket Factors in Black–White Wage Differences." *Journal of Political Economy* 104 (5): 869–95.

Neumark, David. 1988. "Employers' Discriminatory Behavior and the Estimation of Wage Discrimination." *Journal of Human Resources* 23 (3): 279–95.

Neumark, David. 1996. "Sex Discrimination in Restaurant Hiring: An Audit Study." *Quarterly Journal of Economics* 111 (3): 915–41.

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

Neumark, David, Ian Burn, and Patrick Button. Forthcoming. "Is It Harder for Older Workers to Find Jobs? New and Improved Evidence from a Field Experiment." *Journal of Political Economy*.

Neumark, David, and Patrick Button. 2014. "Did Age Discrimination Protections Help Older Workers Weather the Great Recession?" *Journal of Policy Analysis and Management* 33 (3): 566–601.

Neumark, David, and Judith Rich. Forthcoming. "Do Field Experiments on Labor and Housing Markets

Overstate Discrimination? A Re-examination of the Evidence." *Industrial and Labor Relations Review*.

Neumark, David, Joanne Song, and Patrick Button. 2017. "Does Protecting Older Workers from Discrimination Make It Harder to Get Hired? Evidence from Disability Discrimination Laws." *Research on Aging* 39 (1): 29–63.

Neumark, David, and Wendy A. Stock. 2006. "The Labor Market Effects of Sex and Race Discrimination Laws." *Economic Inquiry* 44 (3): 385–419.

Niederle, Muriel, and Lise Vesterlund. 2007. "Do Women Shy Away from Competition? Do Men Compete Too Much." *Quarterly Journal of Economics* 122 (3): 1067–101.

Norton, Michael I., Samuel R. Sommers, Evan P. Apfellbaum, Natassia Pura, and Dan Ariely. 2006. "Color Blindness and Interracial Interaction: Playing the Political." *Psychological Science* 17 (11): 349–53.

Norton, Michael I., Joseph A. Vandello, and John M. Darley. 2004. "Casuistry and Social Category Bias." *Journal of Personality and Social Psychology* 87 (6): 817–31.

Nunley, John M., Adam Pugh, Nicholas Romero, and R. Alan Seals. 2017. "The Effects of Unemployment and Underemployment on Employment Opportunities: Results from a Correspondence Audit of the Labor Market for College Graduates." *ILR Review* 70 (3): 642–69.

Oaxaca, Ronald. 1973. "Male–Female Wage Differentials in Urban Labor Markets." *International Economic Review* 14 (3): 693–709.

Olian, Judy D., Donald P. Schwab, and Yitchak Haberfeld. 1988. "The Impact of Applicant Gender Compared to Qualifications on Hiring Recommendations: A Meta-analysis of Experimental Studies." *Organizational Behavior and Human Decision Processes* 41 (2): 180–95.

Oreopoulos, Philip. 2011. "Why Do Skilled Immigrants Struggle in the Labor Market? A Field Experiment with Thirteen Thousand Resumes." *American Economic Journal: Economic Policy* 3 (4): 148–71.

Oswald, Frederick L., Gregory Mitchell, Hart Blanton, James Jaccard, and Philip E. Tetlock. 2013. "Predicting Ethnic and Racial Discrimination: A Meta-analysis of IAT Criterion Studies." *Journal of Personality and Social Psychology* 105 (2): 171–92.

Pager, Devah. 2003. "The Mark of a Criminal Record." *American Journal of Sociology* 108 (5): 937–75.

Pager, Devah. 2007. "The Use of Field Experiments for Studies of Employment Discrimination: Contributions, Critiques, and Directions for the Future." *Annals of the American Academy of Political and Social Science* 609: 104–33.

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.

Parsons, Christopher A., Johan Sulaeman, Michael C. Yates, and Daniel S. Hamermesh. 2011. "Strike Three: Discrimination, Incentives, and Evaluation." *American Economic Review* 101 (4): 1410–35.

Petit, Pascale. 2007. "The Effects of Age and Family Constraints on Gender Hiring Discrimination: A Field Experiment in the French Financial Sector." *Labour Economics* 14 (3): 371–91.

Phillips, David C. Forthcoming. "Do Comparisons of Fictional Applicants Measure Discrimination When Search Externalities Are Present? Evidence from Existing Experiments." *Economic Journal*.

Pingitore, Regina, Bernard L. Dugoni, R. Scott Tindale, and Bonnie Spring. 1994. "Bias against Overweight Job Applicants in a Simulated Employment Interview." *Journal of Applied Psychology* 79 (6): 909–17.

Polinko, Natale K., and Paula M. Popovich. 2001. "Evil Thoughts but Angelic Actions: Responses to Overweight Job Applicants." *Journal of Applied Social Psychology* 31 (5): 905–24.

Price, Joseph, and Justin Wolfers. 2010. "Racial Discrimination among NBA Referees." *Quarterly Journal of Economics* 125 (4): 1859–87.

Quillian, Lincoln, Devah Pager, Ole Hexel, and Arnfinn H. Midtboen. 2017. "Meta-analysis of Field Experiments Shows No Change in Racial Discrimination in Hiring Over Time." *Proceedings of the National Academy of Sciences* 114 (41): 10870–75.

Ravaud, Jean François, Béatrice Madiot, and Isabelle Ville. 1992. "Discrimination towards Disabled People Seeking Employment." *Social Science & Medicine* 35 (8): 951–58.

Reuben, Ernesto, Paola Sapienza, and Luigi Zingales. 2014. "How Stereotypes Impair Women's Careers in Science." *Proceedings of the National Academy of Sciences* 111 (12): 4403–08.

Riach, Peter A. 2015. "A Field Experiment Investigating Age Discrimination in Four European Labour Markets." *International Review of Applied Economics* 29 (5): 608–19.

Riach, Peter A., and Judith Rich. 2002. "Field Experiments of Discrimination in the Market Place." *Economic Journal* 112 (483): F480–518.

Riach, Peter A., and Judith Rich. 2006. "An Experimental Investigation of Age Discrimination in the French Labour Market." Institute for the Study of Labor Discussion Paper 2522.

Riach, Peter A., and Judith Rich. 2010. "An Experimental Investigation of Age Discrimination in the English Labor Market." *Annals of Economics and Statistics* 99–100: 169–85.

Rich, Judith. 2014. "What Do Field Experiments of Discrimination in Markets Tell Us? A Meta Analysis of Studies Conducted since 2000." Institute for the Study of Labor Discussion Paper 8584.

Riggio, Ronald E., and Barbara Throckmorton. 1988. "The Relative Effects of Verbal and Nonverbal Behavior, Appearance, and Social Skills on Evaluations Made in Hiring Interviews." *Journal of Applied Social Psychology* 18 (4): 331–48.

Rooth, Dan-Olof. 2009. "Obesity, Attractiveness, and Differential Treatment in Hiring: A Field Experiment." *Journal of Human Resources* 44 (3): 710–35.

Rooth, Dan-Olof. 2010. "Automatic Associations and

Discrimination in Hiring: Real World Evidence." *Labour Economics* 17 (3): 523–34.
Rosen, Benson, and Thomas H. Jerdee. 1974. "Influence of Sex Role Stereotypes on Personnel Decisions." *Journal of Applied Psychology* 59 (1): 9–14.
Rosen, Benson, and Thomas H. Jerdee. 1976. "The Influence of Age Stereotypes on Managerial Decisions." *Journal of Applied Psychology* 61 (4): 428–32.
Rosen, Benson, and Thomas H. Jerdee. 1977. "Too Old or Not Too Old?" *Harvard Business Review* 55 (6): 97–106.
Rudolph, Cort W., Charles L. Wells, Marcus D. Weller, and Boris B. Baltes. 2009. "A Meta-analysis of Empirical Studies of Weight-Based Bias in the Workplace." *Journal of Vocational Behavior* 74 (1): 1–10.
Ruffle, Bradley J., and Ze'ev Shtudiner. 2015. "Are Good-Looking People More Employable?" *Management Science* 61 (8): 1760–76.
Schwartz, Richard D., and Jerome H. Skolnick. 1962. "Two Studies of Legal Stigma." *Social Problems* 10 (2): 133–42.
Siddique, Zahra. 2011. "Evidence on Caste Based Discrimination." *Labour Economics* 18 (Supplement 1): S146–59.
Singer, M. S., and Christine Sewell. 1989. "Applicant Age and Selection Interview Decisions: Effect of Information Exposure on Age Discrimination in Personnel Selection." *Personnel Psychology* 42 (1): 135–54.
Solnick, S. J. 2001. "Gender Differences in the Ultimatum Game." *Economic Inquiry* 39 (2): 189–200.
Steffens, Melanie C. 2004. "Is the Implicit Association Test Immune to Faking?" *Experimental Psychology* 51 (3): 165–79.
Tilcsik, András. 2011. "Price and Prejudice: Employment Discrimination against Openly Gay Men in the United States." *American Journal of Sociology* 117 (2): 586–626.
Tosi, Henry L., and Steven W. Einbender. 1985. "The Effects of the Type and Amount of Information in Sex Discrimination Research: A Meta-analysis." *Academy of Management Journal* 28 (3): 712–23.
Turner, Margery Austin, Michael Fix, and Raymond J. Struyk. 1991. *Opportunities Denied, Opportunities Diminished: Racial Discrimination in Hiring*. Washington, DC: Urban Institute Press.
US Department of Labor. 1965. *The Older American Worker: Age Discrimination in Employment*. Washington, DC: US Government Printing Office.
Weichselbaumer, Doris. 2003. "Sexual Orientation Discrimination in Hiring." *Labour Economics* 10 (6): 629–42.
Weichselbaumer, Doris. 2004. "Is It Sex or Personality? The Impact of Sex Stereotypes on Discrimination in Applicant Selection." *Eastern Economic Journal* 30 (2): 159–86.
Weichselbaumer, Doris. 2015a. "Beyond the Veil: Discrimination against Female Migrants Wearing a Headscarf in Germany." Unpublished.
Weichselbaumer, Doris. 2015b. "Testing for Discrimination against Lesbians of Different Marital Status: A Field Experiment." *Industrial Relations* 54 (1): 131–61.
Weiss, Elizabeth M., and Todd J. Maurer. 2004. "Age Discrimination in Personnel Decisions: A Reexamination." *Journal of Applied Social Psychology* 34 (8): 1551–62.
Word, Carl O., Mark P. Zanna, and Joel Cooper. 1974. "The Nonverbal Mediation of Self-Fulfilling Prophecies in Interracial Interaction." *Journal of Experimental Social Psychology* 10 (2): 109–20.
Zhou, Xiangyi, Jie Zhang, and Xuetao Song. 2013. "Gender Discrimination in Hiring: Evidence from 19,130 Resumes in China." Unpublished.
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.
Zussman, Asaf. 2013. "Ethnic Discrimination: Lessons from the Israeli Online Market for Used Cars." *Economic Journal* 123 (572): F433–68.