NBER WORKING PAPER SERIES

FIELD EXPERIMENTS ON DISCRIMINATION

Marianne Bertrand
Esther Duflo

Working Paper 22014
http://www.nber.org/papers/w22014

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
February 2016

Laura Stilwell and Jan Zilinsky provided excellent research assistance. We thank Abhijit Banerjee for comments. We are particularly grateful to Betsy Levy Paluck, our discussant, for her detailed and thoughtful review of an earlier draft. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Field Experiments on Discrimination
Marianne Bertrand and Esther Duflo
NBER Working Paper No. 22014
February 2016
JEL No. J0,J01,J1,J15,J16,J7,J71

ABSTRACT

This article reviews the existing field experimentation literature on the prevalence of discrimination, the consequences of such discrimination, and possible approaches to undermine it. We highlight key gaps in the literature and ripe opportunities for future field work. Section 1 reviews the various experimental methods that have been employed to measure the prevalence of discrimination, most notably audit and correspondence studies; it also describes several other measurement tools commonly used in lab-based work that deserve greater consideration in field research. Section 2 provides an overview of the literature on the costs of being stereotyped or discriminated against, with a focus on self-expectancy effects and self-fulfilling prophecies; section 2 also discusses the thin field-based literature on the consequences of limited diversity in organizations and groups. The final section of the paper, Section 3, reviews the evidence for policies and interventions aimed at weakening discrimination, covering role model and intergroup contact effects, as well as socio-cognitive and technological de-biasing strategies.

Marianne Bertrand
Booth School of Business
University of Chicago
5807 South Woodlawn Avenue
Chicago, IL 60637
and NBER
marianne.bertrand@chicagobooth.edu

Esther Duflo
Department of Economics, E17-201B
MIT
77 Massachusetts Avenue
Cambridge, MA 02139
and NBER
eduflo@mit.edu

Introduction

Black people are less likely to be employed, more likely to be arrested by the police, and more likely to be incarcerated. Women are very scarce at the top echelon of the corporate, academic and political ladders despite the fact that (in rich countries at least) they get better grades in school and are more likely to graduate from college. While many in the media and public opinion circles argue that discrimination is a key force in driving these patterns, showing that it is actually the case is not simple. Indeed, it has proven elusive to produce convincing evidence of discrimination using standard regression analysis methods and observational data, in the sense in which we define discrimination throughout this chapter: *members of a minority group (women, Blacks, Muslims, immigrants, etc.) are treated differentially (less favorably) than members of a majority group with otherwise identical characteristics in similar circumstances.*

However, over the last couple of decades, a rich literature in economics, sociology, political science and psychology has leveraged experiments (in the lab and in the field) to provide convincing evidence that discrimination, thus defined, indeed exists. We begin this chapter by describing the various experimental methods that have been used to measure discrimination in the field. Overall, this literature offers staggering evidence of pervasive discrimination against minority or under-represented groups all around the world. We summarize this research and discuss some of its key limitations.

If discrimination is as pervasive as the evidence suggests, what do existing theories tell us about the costs to minority groups and to society overall? The two workhorse models of discrimination in the economics literature give drastically different answers, particularly with respect to the societal consequences. In the first model, developed in Becker (1957) for the context of the labor market, some employers have a distaste for hiring members of the minority group. They may indulge this distaste by refusing to hire, say, Blacks or, if they do hire them, paying them less than other employees for the same level of productivity. If the fraction of discriminating employers in the economy is sufficiently large, a wage differential will emerge in equilibrium between otherwise identically productive minority and majority employees and this wage differential will be a reflection of the distaste parameter of the marginal employer for minority workers (Becker, 1957; Charles and Guryan, 2008). By electing to not hire minority workers, infra-margin racist employers will experience lower profits. In fact, if the conditions of perfect competition were

satisfied, discriminating employers would be wiped away and taste-based discrimination would disappear.[1]

This "taste-based" explanation for discrimination stands in contrast with what many economists would view as a more disciplined explanation, which does not involve an ad-hoc (even if intuitive) addition to the utility function (animus towards certain groups) to help rationalize a puzzling behavior. In a "statistical discrimination" model (Phelps, 1972; Arrow, 1973; Aigner and Cain, 1977), the differential treatment of members of the minority group is due to imperfect information, and discrimination is the result of a signal extraction problem. As a profit-maximizing prospective employer, renter, or car salesman, tries to infer the characteristics of a person that are relevant to the market transaction they are considering to complete with that person, they use all the information available to them. When the person-specific information is limited, group-specific membership may provide additional valuable information about expected productivity. For example, again using the labor market scenario, it may be known to employers that minority applicants are on average less productive than majority applicants. In this case, an employer who sees two applicants with similar noisy but unbiased signals of productivity should rationally favor the majority applicant to the minority one as her expected productivity is higher. While expected productivity will equal true productivity on average within each group, statistical discrimination will result in some minority workers being treated less favorably than non-minority workers of the same true productivity, i.e. will result in discrimination as defined above. In the extreme case where individual signals of productivity are totally uninformative, an employer may rationally decide to make offers only to Whites if the mean productivity among Blacks does not exceed the required threshold.

While taste-based discrimination is clearly inefficient (simply consider how it constrains the allocation of talent), statistical discrimination is theoretically efficient and hence more easily defendable in ethical terms under the utilitarian argument. Moreover, statistical discrimination can also be argued to be "fair" in that it treats identical people with the same expected productivity (even if not with the same actual productivity) and is not motivated by animus. In fact, many economists would most likely support allowing statistical discrimination as a good policy, even where it is now illegal (as it is, for example, in the US labor market and real estate market

[1] Refusing to hire black people could be efficient if the employer knows he cannot work well with them due to his animus, but this does not take away from the fact that businesses that do this should not survive.

contexts).

Unfortunately, as we discuss below, while field experiments have been overall successful at documenting that discrimination exists, they have (with a few exceptions) struggled with linking the patterns of discrimination to a specific theory.

Meanwhile, psychologists have made considerable progress in their own understanding of the roots of discrimination, on a largely parallel track. The theories they have advanced and the (mainly lab) experiments they have conducted have been helpful in better nailing the micro-foundations of discrimination. We believe this body of work blurs the sharp line economists tend to draw between taste-based and statistical explanations.

Psychologists' work on discrimination is embedded in an immense literature that attempts to understand the roots of prejudice, widely characterized as negative evaluation of others made on the basis of their group membership. This literature has looked for the micro-foundations of such negative evaluation in a wide variety of areas, including personality development, socialization, social cognition, evolutionary psychology and neuroscience.

Early scholarship in psychology viewed prejudice as a form of abnormal thinking and equated it to a psychopathology (think Adolf Hitler) that could be treated by addressing the personality disorders of subset of the population that was "diseased." It was only in the second half of the twentieth century that that the prevalent view of prejudice among psychologists became rooted in normal thinking processes (Dovidio, 2001), with socialization and social norms being viewed as dominant drivers. Influential work by Tajfel (Tajfel, 1970; Tajfel and Turner, 1979) demonstrated the key role social identity plays in the process underlying prejudice. Experimental evidence has shown that the assignment of people to groups, even if totally arbitrary ones and even if they do not last, is sufficient to produce favoritism for in-group members and negativity towards out-group members. At the same time, evolutionary psychology has stressed the importance of social differentiation and the delineation of clear group boundaries as a way to achieve the benefits of cooperation between human beings without the risk of excessive costs, with group membership and group identity emerging as a form of contingent altruism (Brewer, 1981). While in-group love might not necessarily imply out-group hate, the same factors that make allegiance with group members important provide grounds for antagonism and distrust of outsiders.

In addition, more recent advances in the psychology literature have demonstrated the existence of unconscious, unintentional forms of bias. Modern social psychologists believe that

attitudes can occur in implicit modes and that people can behave in ways that are unrelated or even sometimes opposed to their explicit views or self-interests (Banaji and Greenwald, 1995; Bertrand et al., 2005; Dovidio et al., 1998a,b; Greenwald and Banaji, 1995). Neuroscience studies have shown that different regions of the brain are activated in conscious versus unconscious processing, suggesting that unconscious processes are distinct mental activities. For example, the unconscious processing of black faces has been associated with activations of area of the brain associated with emotions and fear while the conscious processing of the same faces increases brain activity in areas related to control and regulation. Implicit biases are more likely to drive behavior under conditions of ambiguity, high time pressures and cognitive loads, or inattentiveness to the task.

Both of these dominant views of prejudice in the psychology literature – as an evolutionary phenomenon making group membership an important component of one's social identity or as an unconscious automatic negative association triggered by exposure to out-group members – could serve as micro-foundations to what the more reduced form "animus-based" models economists have worked with. More importantly, these psychological models make clear that the limited information and decision-making model that drives statistical discrimination might be itself endogenous to conscious or unconscious prejudice against the out-group members. If a social need to positively associate with one's own group also makes the out-group members feel more distant and unknowable (Brewer, 1988), an employer may not invest in collecting information on an out-group member, or decide that the individual signals of productivity for minority group members are totally uninformative, resulting in all minority group members being equally treated as un-hirable. Limited de-facto contact between in-group and out-group members will imply that majority employees or coworkers will be fairly ignorant about the quality of minorities; this would mean that employing, electing, or renting to them may seem riskier which, in the presence of risk aversion, will also trigger more statistical discrimination (Aigner and Cain, 1977). Unconscious bias may influence the specific criteria or formulae that are used to assess expected productivity (Uhlmann and Cohen, 2005): for example, the sense of danger that is implicitly triggered by seeing a black face or reading a black name on a résumé may lead an employer to put too great a weight on docility as a work quality than would be warranted for maximum productivity. Recently, the emphasis on "fit" between a prospective employee and the company as a hiring criterion in technology jobs has raised the spectrum of a new form of subtle discrimination.

Similarly, unconscious stereotypes may influence our judgment of the inputs into productivity, with the same level of assertiveness being deemed as good for productivity when coming from men but bad when coming from women (Rudman and Glick, 2001).

Perhaps most importantly, whether discrimination is taste-based or statistical, it may ultimately result in genuine difference between groups, through self-fulfilling prophecies. If the stereotypical woman is not good at math, talented girls may become discouraged and ultimately not become good at math. If teachers or employers assume that students of a particular color are less smart, they will invest less in them. Thus, discrimination, whether it is taste-based or statistical, can create or exacerbate existing differences between groups. Discrimination that starts as taste-based and inefficient can easily morph into the more "justifiable" form. "Valid" stereotypes today could be the product of ambient animus, very much complicating the division between the different theories of discrimination.

The chapter proceeds as follows. Section 1 is devoted to the various experimental methods that have been used in the field to measure discrimination, in particular audit and correspondence studies. Audit studies send out individuals who are matched in all observable characteristics except for the one in question (race, criminal record, etc.) to apply for jobs or make purchases, then researchers analyze the responses. Correspondence studies – which represent by far the largest share of field experiments on discrimination – do the same but control for more variables by creating fictitious applicants (often for jobs or apartments) who correspond via mail. We summarize the findings of this body of work (which clearly demonstrate the pervasiveness of discrimination) and discuss its key limitations.

In this section of the chapter, we also discuss a few alternative methods to measure discrimination, many of them having developed in the psychology literature for use in the lab: Implicit Association Tests, Goldberg Paradigm experiments and List Randomization – as well as measures of willingness to pay to interact with minority group members. We argue that these alternative methods deserve more consideration by economists interested in measures of discrimination for their field research.

Section 2 reviews the work that addresses the costs of being discriminated against, or stereotyped. In particular, we review the experimental work that has studied how the threat of being viewed through the lens of a negative stereotype can have a direct negative effect on performance. We also review the experimental literature on expectancy effects, the goal of which has been to

understand how stereotypes and biases against minority groups may end up being self-fulfilling.

We round up the second part of the chapter by reviewing what is a surprisingly thin field-based literature on the costs (and benefits) of the limited diversity in organizations and groups that directly result from discrimination. This allows us to discuss field work that has considered the consequences of discrimination not just from the perspective of the group that is discriminated against, but also from the perspective of society as a whole.

The third and final section of this chapter, Section 3, is related to the review of various interventions and policies that have been proposed to undo or weaken discrimination. This section covers topics such as the impact of role models, how contact and exposure to the minority groups may change prejudice, as well as a large psychology literature on both socio-cognitive and technological de-biasing strategies. We argue that there is a lot of promising future research that is "ripe for the picking" in this area, given the large amount of theoretical and lab-based work that has not yet been taken to the field.

1 Measuring Discrimination in the Field

Earlier research on discrimination focused on individual-level outcome regressions, with discrimination estimated from the "minority" differential that remains unexplained after including as many proxies as possible for productivity.[2]

The limitations of this approach are well-known. The interpretation of the estimated "minority" coefficient is problematic due to omitted variables bias. Specifically, results of a regression analysis might suggest differential treatment by race or gender even if the decision-maker (say, an employer) never used group membership in her decision of how much to pay an employee. However, it could be the case that race or gender is correlated with other proxies for productivity that are unobservable to the researcher but observed by the employer. It is therefore impossible to conclude that the employer used group membership in her decision-making process using this method.

The traditional answer has been to saturate the regression with as many productivity-relevant, individual-level characteristics as are available. But, of course, ensuring that the researcher observes all that the decision-maker observes is a hopeless task.

[2]For a review of this earlier literature on the narrower topic of labor market discrimination, see Chapter 48 in Altonji and Blank (1999).

Moreover, adding more and more controls to a regression could ultimately obscure the interpretation of the evidence. Consider the labor market context: minority workers might be best-responding to the discrimination they know to exist and could have simply sorted into industries where there is no or limited discrimination. Hence, finding no racial gap in earnings after controlling for industry or employer fixed effects in a regression may indicate that there is no discrimination at the margin, which is very different from no discrimination on average.

Also, as pointed out in Guryan and Charles (2013), the variables the researcher controls for might themselves be affected by discrimination. That is, disadvantaged groups may not have access to high-quality schools because of discrimination, yet they might, given their low human capital accumulation, be paid the "fair market wage." While one might still be tempted to conclude from this that there is no discrimination in the labor market but instead discrimination in the education market, that might not be right if the minority group's expectations about labor market discrimination drive their educational decision. In other words, minority group members may decide to underinvest in education if they expect that they will not be able to obtain labor market returns for this education.

Audit and correspondence methodologies were developed to address these core limitations of the regression approach to measuring discrimination. We review below both types of studies, discuss the extent to which they address these limitations of the regression approach, and also consider new issues they create.

1.1 Audit Studies

In the best-known collection of audit studies exploring the extent of discrimination, Fix and Struyk (1993), describe the method as follows:

> Two individuals (auditors or testers) are matched for all relevant personal characteristics other than the one that is presumed to lead to discrimination, e.g. race, ethnicity, gender. They then apply for a job, a housing unit, or a mortgage, or begin to negotiate for a good or service. The results they achieve and the treatment they receive in the transaction are closely observed, documented, and analyzed to determine if the outcomes reveal patterns of differential treatment on the basis of the trait studied and/or protected by anti-discrimination laws...

Discrimination is said to have been detected when "auditors in the protected class are systematically treated worse than their teammates" (Yinger, 1998).[3]

A well-known early example of the audit method is offered by Ayres and Siegelman (1995). In this study, pairs of testers (one of whom was always a white male) were trained to bargain uniformly and then were sent to negotiate for the purchase of a new automobile at randomly selected Chicago-area dealerships. Thirty-eight testers bargained for 306 cars at 153 dealerships. Testers were chosen to have average attractiveness, and both testers in a pair bargained for the same model of car, at the same dealership, usually within a few days of each other. Dealerships were selected randomly, testers were randomly assigned to dealerships, and the choice of which tester in the pair would be the first to enter the dealership was also randomly made. The testers bargained at different dealerships for a total of nine car models, following a uniform bargaining script that instructed them to focus quickly on one particular car and start negotiating over it. Testers were further instructed to tell dealers at the beginning of the bargaining that they could provide their own financing for the car. In spite of the identical approach to bargaining, Ayres and Siegelman (1995) find that white males are quoted lower prices than white women and Blacks (men or women). While ancillary evidence suggests that the dealerships' disparate treatment of women and Blacks may be caused by dealers' statistical inferences about consumers' reservation prices, the data do not strongly support any single theory of discrimination.

Another well-known audit study of the labor market is Neumark, Bank, and Van Nort (1996), which investigates the role of sex discrimination in vertical segregation among waiters and waitresses. Specifically, two male and two female college students were sent to apply in person for jobs as waiters and waitresses at 65 restaurants in Philadelphia. The restaurants were divided into high-, medium-, and low-price categories, with the goal of estimating sex differences in the receipt of job offers in each price category. The study was designed so that a male and female pair applied for a job at each restaurant, and so that the male and female candidates were on average identical. The findings are consistent with discrimination against women in high-price restaurants and discrimination in women's favor in low-price restaurants. Of the thirteen job offers from high-price restaurants, eleven were made to men. In contrast, of the ten job offers from low-price restaurants, eight were made to women. In addition, information gathered from

[3]Results from the earliest audit studies can be found in Newman (1978), McIntyre, Moberg, and Posner (1980), Galster (1990), Yinger (1986), Cross, Kenney, Mell, and Zimmerman (1990), James and DelCastillo (1991), Turner, Fix, and Struyk (1991), and Fix and Struyk (1993).

restaurants included in the study suggests that earnings are substantially higher in high-price restaurants, meaning that the apparent hiring discrimination has implications for gender-based differences in earnings among waitpersons. Results are interpreted as consistent both with employer discrimination and customer discrimination.

Another interesting application of the audit method is Pager (2003) who matched pairs of individuals applying for entry-level positions, and probed the impact of a criminal record, conditional on race. The author employed two black testers who formed a team, and another pair of white testers. Within each team, one auditor was "assigned" a criminal record (this assignment was random and rotating – that is, each tester played the role of an ex-convict at some point).[4] In total, 350 employers were audited. The effect of the criminal record was both statistically significant and meaningful in magnitude: 17 percent of attempts with Whites who had a supposed criminal record received a call-back, compared to 34 percent of tries with Whites who said they had no criminal record. That is, an equally qualified white candidate was about half as likely to receive a call-back if he was believed to be an ex-convict. For black applicants, the effect was notably larger: 5 percent of attempts with Blacks who were supposedly ex-convicts received a call-back, compared to 14 percent of applications with Blacks that had no record, meaning an equally qualified black candidate was about one-third as likely to receive a call-back if he had a criminal record. Furthermore, these estimates show that a black applicant without a criminal record was about as likely to receive a call-back as a white applicant *with* a criminal record.

Most audit studies do not explicitly test which theory of discrimination has most explanatory power, even if they often informally discuss what forms of discrimination might or might not be consistent with the observed patterns in the data. A notable exception is List (2004) who recruited buyers and sellers at a sports cards market and documented that minority buyers receive lower offers when they bargain for a collectible card. One finding of List (2004) is that in this context lack of information – and the expectation that minorities are inexperienced – drives discriminatory behavior. Experienced dealers discriminate more. Among experienced buyers, final offers to minorities are similar to offers received by white men; but minorities require more time to achieve this outcome. Moreover, List tries to rule out taste-based explanations for the

[4]Pager argues that "[b]y varying which member of the pair presented himself as having a criminal record, unobserved differences within the pairs of applicants were effectively controlled for."

data by combining the field data with results from a dictator game conducted in the lab with these card dealers. He finds that non-white males receive roughly as many positive allocations in this game as white males and interprets this pattern as evidence for the absence of taste for discrimination. Of course, while a laboratory experiment is a useful complement to the field study, the behavior of dealers in the dictator game, on its own, does not prove that taste-based discrimination is absent during the actual market transactions.

1.1.1 Limitations of Audit Studies

Many of the weaknesses of audit studies have been discussed in Heckman and Siegelman (1993) and Heckman (1998). First, these studies require that both members of the auditor pair be identical in all dimensions that might affect productivity in employers' eyes, except for the trait that is being manipulated. To accomplish this, researchers typically match auditors on several characteristics (height, weight, age, dialect, dressing style, hairdo) and train them for several days to coordinate interviewing styles. Yet, critics note that this is unlikely to erase the numerous differences that exist between the auditors in a pair.

Another weakness of the audit studies is that they are not double-blind: auditors know the purpose of the study. As Turner, Fix, and Struyk (1991), note: "The first day of training also included an introduction to employment discrimination, equal employment opportunity, and a review of project design and methodology." This may generate conscious or subconscious motives among auditors to generate data consistent or inconsistent with their beliefs about race or gender issues. As psychologists have documented, these demand effects can be quite strong. It is very difficult to insure that auditors will not want to do "a good job." Even a vague belief by auditors that employers treat minorities differently can result in measured differences in treatment. The possibility of such a demand effect is further magnified by the fact that auditors are not in fact seeking jobs (or trying to buy a car for themselves) and are therefore more free to let their beliefs affect the bargaining or interview process.

1.2 Correspondence Studies

Correspondence studies have been developed to address some of the more obvious weaknesses of the audit method. Rather than relying on real auditors or testers that physically meet with a potential employer or potential landlord, correspondence studies rely on fictitious applicants.

Specifically, in response to a job or rental advertisement, the researcher sends (many) pairs of résumés or letters of interest, one of which is assigned the perceived minority trait. Discrimination is estimated by comparing the outcomes for the fictitious applicants with and without the perceived minority trait. The most common (but not the only) way to manipulate the perceived minority trait has been through the names of the applicants (e.g. female names, African-American names, Arabic Names, etc). Outcomes studied in a correspondence study have been mainly, but not exclusively, limited to measuring call-backs by employers or landlords in response to the mailed or emailed fictitious application.[5]

The correspondence method presents several advantages over the audit method. First, because it relies on résumés or applications by fictitious people and not real people, one can be sure to generate strict comparability across groups for *all* information that is seen by the employers or landlords. This guarantees that any observed differences are caused solely by the minority trait manipulation. Second, the use of paper applications insulates from demand effects. Finally, because of the relatively low marginal cost, one can send out a large number of applications. Besides providing more precise estimates, the larger sample size also allows researchers to examine the nature of the differential treatment from many more angles, and hence promises to link it more closely to specific theories of discrimination.[6]

Although Guryan and Charles (2013) call correspondence tests a "significant methodological advance," and a review of discrimination in the marketplace published about 15 years ago discussed only observational and audit studies (Yinger, 1998), the method is actually not that new. Fictitious applications and résumés have been sent to employers in order to uncover racial or religious discrimination nearly half a century ago.[7] However, the number of correspondence studies in economics has greatly increased following Bertrand and Mullainathan (2004), who study race discrimination in the labor market by sending fictitious résumés in response to help-wanted ads in Boston and Chicago newspapers. To manipulate perceived race, they randomly assigned very white-sounding names (such as Emily Walsh or Greg Baker) to half the résumés and very African-American-sounding names (such as Lakisha Washington or Jamal Jones) to the other

[5] See Sections 1.3.2 and 1.3.3 for cases of different approaches.

[6] We discuss in Section 1.5 other weaknesses that are shared by the correspondence studies, as well as added weaknesses of the correspondence method compared to the audit method.

[7] See Jowell and Prescott-Clarke (1970), Jolson (1974), Hubbuck and Carter (1980), Brown and Gay (1985), and Riach and Rich (1991) for early studies. One caveat is that some of these studies fail to fully match skills between minority and non-minority résumés.

half. In total, they responded to over 1,300 employment ads in the sales, administrative support, clerical, and customer services job categories and sent out nearly 5,000 résumés. They find that white names receive 50 percent more call-backs for interviews.

1.2.1 Correspondence studies in the labor market

The main results of labor market correspondence tests are reviewed in Table 1.

As is clear from Table 1, labor market correspondence studies have by now been carried in many countries around the world, and have focused on a variety of perceived traits that can be randomized on a résumé. Below, we review some of these studies in more detail, focusing in particular on those that have attempted to go beyond simply documenting whether or not differential treatment occurs based on the manipulated traits, and move toward understanding which theory may best fit the patterns in the data. However, one of our bottom lines will be that, unfortunately, the studies have tended to be fairly close replications of Bertrand and Mullainathan (2004) for different populations or contexts. With a few exceptions, the literature has failed to push the correspondence methodology to design approaches to more formally test for various theories of why differential treatment is taking place.

Race, ethnicity Studies of labor market discrimination based on race and ethnic background have been by far the most popular application of the correspondence method to date. While publication bias is always a concern, the results of correspondence studies where the trait of interest is race or ethnicity offer overwhelming evidence of discrimination in the labor market against racial and ethnic minorities. Evidence has been accumulated from nearly all continents: Latin America (e.g. Galarza and Yamada (2014) compare Whites to indigenous applicants in Peru), Asia (e.g. Maurer-Fazio (2012) compares Han, Mongolian and Tibetan applicants in China), Australia (e.g. where Booth, Leigh, and Varganova (2011) compare Whites to Chinese applicants), Europe (e.g. Baert et al. (2013) compare immigrants to non-immigrants in Belgium), Ireland (e.g. where McGinnity et al. (2009) compare candidates with Irish names to candidates with distinctively non-Irish names), etc.

Attempts to adapt the correspondence method to learn more about which theory of discrimination best fits the patterns in the data have been mainly focused on trying to provide corroborative evidence for (or against) statistical discrimination. The most common approach

Table 1: Labor market correspondence studies

Paper	Country	CVs / apps	Vacancies	Effect (Call-back ratio)	Theory
Galarza and Yamada (2014) Trait: Ethnicity; Attractiveness	Peru	4,820	1,205	White-to-indigenous ratio: 1.8 Low attractiveness hurts white females	No
Eriksson and Rooth (2014) Trait: Unemployment duration	Sweden	8,466	-	Employed to long-term unemployed: 1.25	No
Blommaert, Coenders, and van Tubergen (2014) Trait: Arabic name	Netherlands	636	-	Dutch-to-foreign: 1.62 (unconditional ratio). No difference, if views held fixed	No
Nunley, Pugh, Romero, and Seals (2014) Trait: Race	US	9,396	-	White-to-black: 1.18 (unconditional)	Inconsistent with statistical discrimination, consistent with taste-based discrimination
Ghayad (2013) Trait: Unemployment duration	US	3360	600	Employed-to-unemployed: 1.47	No
Bartoš, Bauer, Chytilová, and Matějka (2013) Trait: Ethnicity (Roma, Asian, Turkish)	Czech Rep. and Germany	274 (Czech R.) 745 (Ger.)	-	Czech-to-Vietnamese: 1.34 Lower requests for CVs if candidate is Turkish	Consistent with attention discrimination
Wright, Wallace, Bailey, and Hyde (2013) Trait: Religion / ethnicity	US	6,400	1,600	White-to-Muslim: 1.58	Consistent with theoretical models of secularization and cultural distate theory
Kroft, Lange, and Notowidigdo (2013) Trait: Unemployment duration	US (largest 100 MSAs)	12054	3,040	1 log point change in unemployment duration: 4.7 percentage points lower call-back probability	No
Baert, Cockx, Gheyle, and Vandamme (2013) Trait: Nationality (Turkish-sounding name)	Belgium	752	376	Flemish-to-Turkish: 1.03 to 2.05, depending on the occupation	No
Bailey, Wallace, and Wright (2013) Trait: Sexual orientation	US	4,608	1,536	No effect	No
Ahmed, Andersson, and Hammarstedt (2013) Trait: Sexual orientation	Sweden	3,990	-	Heterosexual-to-homosexual (male): 1.14 Heterosexual-to-homosexual (female): 1.22	No
Acquisti and Fong (2013) Traits:Sexual orientation and religion	US	4183	-	Christian-to-Muslim: 1.16	No
Patacchini, Ragusa, and Zenou (2012) Traits: Sexual orientation and attractiveness	Italy	2,320	-	Heterosexual-to-Homosexual: 1.38	No
Kaas and Manger (2012) Trait: Immigrant (race/ethnicity)	Germany	1,056	528	German-to-Turkish: 1.29 (if no reference letter is included)	Consistent with statistical discrimination
Maurer-Fazio (2012) Trait:Ethnicity	China	21,592	10,796	Han-to-Mongolian: 1.36 Han-to-Tibetan: 2.21	No

...Continued from previous page

Paper	Country	CVs / apps	Vacancies	Effect (Call-back ratio)	Theory
Jacquemet and Yannelis (2012) TRAIT: Race / Nationality	US	330	990	English-to-foreign names: 1.41 English-to-Black names: 1.46	Consistent with patterns of ethnic homophily
Ahmed, Andersson, and Hammarstedt (2012) TRAIT: Age	Sweden	466	-	31 year old-to-46 year old: 3.23	No
Oreopoulos (2011) TRAIT: Nationality (and race)	Canada	12910	3225	English name-to-Immigrant: ranged from 1.39 to 2.71 (against Indian Pakistani and Chinese applicants)	No
Carlsson (2011) TRAIT: Gender	Sweden	3,228	1,614	Female-to-Male: 1.07	No
Booth, Leigh, and Varganova (2011) TRAIT: Ethnicity	Australia	Above 4000	-	White-to-Italian: 1.12 White-to-Chinese: 1.68	No
Booth and Leigh (2010) TRAIT: Gender	Australia	3,365	-	Female-to-male: 1.28 (female-dominated professions)	No
Riach and Rich (2010) TRAIT: Age	UK	1,000+	-	2.64 favoring younger candidates	No
Rooth (2009) TRAIT: Attractiveness/Obesity	Sweden	1,970	985	Non-obese/attractive-to-obese/unattractive: ranged from 1.21 to 1.25 (but higher for some occupations)	No
McGinnity, Nelson, Lunn, and Quinn (2009) TRAIT: Nationality / race	Ireland	480	240	1.8, 2.07, 2.44 in favor of Irish and against Asians, Germans and Africans respectively	No
Banerjee, Bertrand, Datta, and Mullainathan (2009) TRAITS: Caste and religion	India	3,160	371	Upper Caste-to-Other Backward Castes: 1.08 (software jobs, insignificant), 1.6 (call-center jobs)	No
Lahey (2008) TRAIT: Age	US	App. 4,000	-	Young-to-older: 1.42	No
Petit (2007) TRAITS: Age, gender, number of children	France	942	157	Ranged from 1.13 to 2.43 against 25-year-old, childless women	No
Bursell (2007) TRAIT: Ethnicity	Sweden	3,552	1,776	Swedish-to-foreign names: 1.82	Inconsistent with statistical discrimination
Bertrand and Mullainathan (2004) TRAIT: Race	US	4,870	1300+	White-to-African-American: 1.5 (1.22 for females in sales jobs)	No
Jolson (1974) TRAIT: Race and religion	US	300	-	White-to-black: 4.2 for selling positions	No

has been to investigate whether the gap in call-backs is responsive to the amount of information provided to employers about the job applicants, as was first done in Bertrand and Mullainathan (2004), in which they studied how credentials affect the racial gap in call-back. In particular, Bertrand and Mullainathan (2004) experimentally varied the quality of the résumé used in response to a given ad. Higher-quality applicants had on average a little more labor market experience and fewer holes in their employment history; they were also more likely to have an e-mail address, have completed some certification degree, possess foreign language skills, or have been awarded some honors. The authors sent four résumés in response to each ad: two higher-quality and two lower-quality ones. They randomly assigned an African-American sounding name to one of the higher- and one of the lower-quality résumés. They find that Whites with higher-quality résumés receive nearly 30 percent more call-backs than Whites with lower-quality résumés. On the other hand, having a higher-quality résumé has a smaller effect for African-Americans. In other words, the gap between Whites and African-Americans widens with résumé quality. While one may have expected improved credentials to alleviate employers' fear that African-American applicants are deficient in some unobservable skills under a statistical discrimination explanation for the overall discrimination, this was not the case in their data. Bertrand and Mullainathan argue that one simple alternative model that may best explain the patterns in their data is some form of lexicographic search by employers:

> Employers receive so many résumés that they may use quick heuristics in reading these résumés. One such heuristic could be to simply read no further when they see an African-American name. Thus they may never see the skills of African-American candidates and this could explain why these skills are not rewarded.

These findings are replicated in Nunley et al. (2014): Blacks received 14 percent fewer call-backs compared to Whites and discrimination was not mitigated when productive characteristics were added to a résumé. However, some studies report results that are more in line with the predictions of statistical discrimination models. Oreopoulos (2011) submitted 12,910 résumés in response to 3,225 job postings in Canada. First, he compares (fictitious) applicants who had a foreign name, but who attended a Canadian (or foreign) university and had work experience in Canada. The call-back rate is 1.39 for Canadian (English-sounding names) versus foreigners if they went to a Canadian university, and 1.43 if they went to a foreign university. However, the

call-back rate falls dramatically if the foreigners' job experience was purely international (2.71 call-back ratio). Moreover, if candidates who had foreign job experience and education had a Chinese last name with an English first name (Allen and Michelle Wang), their prospects on the job market improved. This raises the possibility that a fraction of the "discrimination" is statistical, for example with employers making inference about the candidate's English skills. Perhaps even more striking, Kaas and Manger (2012) sent out 528 pairs of applications in Germany to study the effect of a Turkish-sounding name. The German-to-Turkish call-back rate was 1.29 when no reference letter was included. Discrimination was eliminated when a reference letter, containing indirect information about productivity (such as conscientiousness and agreeableness) was added, which the authors interpret as evidence of consistency with statistical discrimination. It is interesting that such "soft information" presented in the reference letter appears to remove the difference in call-back rates even though "harder information" presented in a résumé (such as employment history or honors) does not in other studies. It would be interesting to probe this contrast further.

Gender There are fewer studies on gender, and discrimination against women at the call-back stage is much less apparent in general. Some studies attempt to show whether the degree (and nature) of discrimination depends on the nature of the profession. Carlsson (2011) sent paired applications for positions of IT professionals, drivers, construction workers, sales assistants, high school teachers, restaurant workers, accountants, cleaners, pre-school teachers, and nurses. Overall, women are called back slightly more often than men; in male dominated professions, males have a slight (insignificant) advantage. Booth and Leigh (2010) focused on female-dominated professions (waitstaff, data-entry, customer service, and sales jobs) and found a call-back of 1.28 in favor of women.

A topic of interest for future work would be to apply the correspondence method to measure the extent to which a bias exists against women with children, or against young women who may have children in the future. To our knowledge only one study, Petit (2007), studies this aspect. In order to shed light on the role of family constraints in gender discrimination, Petit sent résumés for male and female applicants, with or without children, of age 25 or 37. Discrimination against women is detected for young workers in higher skilled positions (in the French finance industry), but not among prime-age workers.

Caste and religion Banerjee et al. (2009) study the role of caste and religion in India's software and call-center sectors. They sent 3160 fictitious résumés with randomly allocated caste-linked surnames in response to 371 job openings in and around Delhi (India) that were advertised in major city papers and online job sites. They find no evidence of discrimination against non-upper-caste (i.e. Scheduled Caste, Scheduled Tribe, and Other Backward Caste) applicants for software jobs. But, in the case of call-center jobs, they do find larger and significant differences between call-back rates for upper-castes and Other Backward Castes (and to a lesser extent Scheduled Castes) in favor of upper-castes. They find no discrimination against Muslims.

The potential impact of religion on job prospects in the US is explored by Wright et al. (2013). Affiliation with a religion was signaled through student activities that were listed on résumés.[8] The control group had no religious identification in his/her résumé. Compared to the control group, Muslim applications were 24 percent less likely to receive at least one contact by either email or phone, and they received 33 percent fewer total contacts than did those from the control group.

Unemployment spells More recently, researchers have applied the correspondence model to better understand patterns of labor market discrimination against the unemployed. In Sweden, Eriksson and Rooth (2014) randomly assigned various characteristics (contemporary unemployment, past unemployment immediately after graduation, past unemployment between jobs, work experience, and number of employers). Long-term unemployment did not harm job candidates' chances, as long as the applicant had subsequent work experience. However, if the applicant was unemployed in the preceding nine months, his or her call-back probability fell by 20 percent.[9] In the US, Ghayad (2013) finds that (current) unemployment spell longer than six months are particularly harmful: the rate of interview requests for résumés with similar firm experience drops 1.13 percentage points for each additional month of non-employment up to six months, and once the candidate experienced six months of unemployment, interview requests fell by an extra 8 percentage points.

[8] It is tricky to signal only religion on a résumé. The manipulation through student activities may reveal more than just religion, an issue we will come back to in Section 1.4.

[9] One caveat, as the authors acknowledge, is that not all employers necessarily view the gaps on the CVs as implying unemployment.

Kroft, Lange, and Notowidigdo (2013) relate these results to the inference problem faced by the prospective employers. The authors replicate the result that longer employment duration reduces call-back rate, but also show that this depends on the labor market conditions. Duration dependence is stronger in tight labor markets, suggesting that employers use the information on the length of unemployment as a signal of productivity, but recognize that the signal is less informative when the labor markets conditions are weak.[10]

Other characteristics: Sexual Orientation and Age Résumé studies are now also being used to try to detect discrimination in a number of less obvious domains.

A literature has tried to estimate discrimination against LGBT candidates, however, most studies have focused only on lesbians and gay men.[11] One of the challenges with estimating discrimination against LGBT candidates is how to provide information that identifies a candidate as a member of that minority, when telling such details are not normally solicited in job applications. In Ahmed, Andersson, and Hammarstedt (2013), which was carried out in Sweden, sexual orientation was indicated by the mention of a "spouse" of either gender in the cover letter, and voluntary work in either an LGBT rights organization (gay identity) or the Swedish red cross (heterosexual identity). Targeted occupations included male-dominated ones (construction worker, motor vehicle driver, sales person, and mechanic worker), female-dominated ones (shop sales assistant, preschool teacher, cleaner, restaurant worker, and nurse), and more neutral ones (high school teacher). The authors find some mild evidence of discrimination (ratio of 1.14), which ultimately could be due to the nature of the signaling (e.g. working in LGBT rights, as opposed to the red cross, may be seen as a political gesture, not just revealing an identity). In Italy, Patacchini, Ragusa, and Zenou (2012) performed a correspondence study that revealed "homosexual preferences" through internships in pro-gay advocacy groups, and finds higher discrimination against gay men (1.38) but not lesbian candidates. In the US, Bailey, Wallace, and Wright (2013) find no evidence of discrimination against gay men or lesbians candidates.

The issue of discrimination by age has also attracted some attention, and several papers (Ahmed, Andersson, and Hammarstedt, 2012; Lahey, 2008; Riach and Rich, 2010) find that

[10]This may also explain the finding in Eriksson and Rooth (2014) (mentioned above) since that particular study was carried out between March and November 2007, i.e. during the global financial crisis.

[11]To the best of our knowledge, no study has been done that specifically looks at discrimination against transgender people using the correspondence method.

younger candidates are generally preferred to older ones. A fundamental issue with this work is that it is hard to argue that age is not necessarily a proxy for productivity. Lahey (2008) tries to control for physical fitness with hobbies (e.g. racquetball is supposed to indicate fitness) but this is ultimately only moderately convincing.

Finally, physical appearance has also been studied: Rooth (2009) studies obesity in the Swedish labor market and Patacchini, Ragusa, and Zenou (2012) investigate the beauty premium in Italy. Using manipulated facial photos to show an otherwise identical candidate as obese, Rooth (2009) shows there is a significantly lower call-back response for obese people: obese men had a six percentage points lower call-back rate, while the call-back rate for obese women was eight percentage points lower. Patacchini, Ragusa, and Zenou (2012) find a small, but significant, beauty premium for "pretty" females (2 percent), however, they do not find a beauty premium for men. Interestingly, the beauty premium disappears for high-skilled attractive women: low-skilled attractive women are more likely to be called back than high-skilled attractive women. On the other hand, Hamermesh and Biddle (1994) do find the existence of a beauty premium in the United States. We discuss the rationale for a beauty premium further in Section 1.9.

1.3 Correspondence Studies in Other Settings

1.3.1 Rental Markets

Correspondence studies in the housing market have very much followed the same approach as those in the labor market. The main findings from the literature are summarized in Table 2.

The rental market studies replicate, in methodology and basic results, those in the labor market. The researchers typically identify rental ads, and send enquiries, manipulating the trait of interest. Discrimination against Arabic names is found in Sweden (Carlsson and Eriksson, 2014; Ahmed and Hammarstedt, 2008; Ahmed et al., 2010). Discrimination against Blacks and other minority ethnicities in the US is found in Ewens, Tomlin, and Wang (2014), Hanson and Hawley (2011) and Carpusor and Loges (2006). Discrimination against immigrants (particularly Muslims) is found in Italy (Baldini and Federici, 2011) and Spain (Bosch, Carnero, and Farré, 2010). Discrimination against LGBT people is found in Ahmed and Hammarstedt (2009).

Another popular variation, parallel to the labor market literature, has been to provide more information (e.g. job, etc.) about some of the applicants. Positive information (e.g. "I do not

Table 2: Rental market papers

Study	Country	Inquiries	Effect	Theory
Carlsson and Eriksson (2014) Trait: Minority status (Arabic name)	Sweden	5,827	Swedish-to-Arabic (females): 1.37 Swedish-to-Arabic (males): 1.62	No
Ewens, Tomlin, and Wang (2014) Trait: Race	US	14,237	White-to-Black: 1.19	Consistent with statistical discrimination, inconsistent with taste-based discrimination
Bartoš, Bauer, Chytilová, and Matějka (2013) Trait: Minority status (Roma or Asian name)	Czech Republic and Germany	1,800	Czech-to-minority: 1.27 (site available), 1.9 (pooled Asian and Roma names)	Consistent with attention discrimination
Hanson and Hawley (2011) Trait: Race	US	9,456	White-to-African American: 1.12 (varied by neighborhood and unit type)	Consistent with statistical discrimination
Baldini and Federici (2011) Trait: Immigrant status; Language ability	Italy	3,676	Italian-to-East European: 1.24 Italian-to-Arab: 1.48	No
Ahmed, Andersson, and Hammarstedt (2010) Trait: Minority status (Arabic name)	Sweden	1,032	Swedish-to-Arab/Muslim: 1.44 (no information), 1.24 (detailed information about the applicant)	No
Bosch, Carnero, and Farré (2010) Trait: Immigrant status	Spain	1,809	Spanish-to-Moroccan: 1.44 (no information), 1.19 (with positive information)	No
Ahmed and Hammarstedt (2009) Trait: Sexual orientation	Sweden	408	Straight-to-gay: 1.27	No
Ahmed and Hammarstedt (2008) Trait: Immigrant (race/ethnicity/religion)	Sweden	1,500	Swedish-to-Arab male: 2.17	No
Carpusor and Loges (2006) Trait: Race / Ethnicity (Arab, African-American)	US (Los Angeles County)	1,115	White-to-Arab: 1.35 White-to-Black: 1.59, conditional on hearing back, 1.98 unconditional	No

smoke and I work full time as an architect") tends to reduce the call-back ratios between white and the minority group, while negative information ("I am a smoker and I have a less than perfect credit score") or small spelling mistakes in the email tend to increase it.

1.3.2 Retail

The expansion of online platforms allows researchers to use the correspondence method to also study discrimination in retail markets. There are currently much fewer such studies, but the door is wide open for more to be performed.

Zussman (2013) studies the mechanisms behind ethnic discrimination in the online market for used cars in Israel. This paper uses an innovative, two-stage approach. First, about 8,000 paired emails are sent to sellers of second-hand cars. An enquiry coming from somebody with a Jewish-sounding name was 22 percent more likely to receive a response than an enquiry emailed by an interested buyer with an Arab-sounding name. Second, a follow-up phone survey was used to elicit sellers' attitudes about minorities to tease out potential mechanisms for this effect. The researchers found that Jewish car sellers who strongly disagree with the statement that "the Arabs in Israel are more likely to cheat than the Jews" do not discriminate against the Arab buyer, while others sellers do. That is, expectations about the quality of the transactions are correlated to the differential (average) treatment of Arabs.

Pope and Sydnor (2011) reports evidence from peer-to-peer lending sites. They find that loan listings with an attached picture of a black individual are 25 to 35 percent less likely to receive funding than those of white individuals with similar credit profiles.

1.3.3 Academia

Milkman, Akinola, and Chugh (2012) ran a field experiment set in academia with a sample of 6,548 professors. Faculty members received e-mails from fictional prospective doctoral students seeking to schedule a meeting either that day or in one week; students' names signaled their race (Caucasian, African-American, Hispanic, Indian, or Chinese) and gender. When the requests were to meet in one week, Caucasian males were granted access to faculty members 26 percent more often than were women and minorities; also, compared with women and minorities, Caucasian males received more and faster responses. However, these patterns were essentially eliminated when prospective students requested a meeting that same day. The authors argue

that their finding of a temporal discrimination effect is consistent with the idea in psychology that subtle contextual shifts can alter patterns of race- and gender-based discrimination (a topic we return to in the last section of this chapter, Section 3.4).

1.4 Beyond the résumés

With the rise of the internet, employers can easily find more information online about a job applicant besides his or her résumé. A few recent studies enrich the correspondence methodology by allowing employers to search for more (and different) information than that which would typically be available in a résumé.

Given the increasing popularity of online social networks, the contribution of Acquisti and Fong (2013) is particularly interesting. They employ the correspondence method by submitting applications to job postings and extend their experiments by creating either personal websites of social networking profiles for the fictitious applicants, which allow employers to gather additional information if they wish to. The additional information that can be gleaned online about the job applicants relates to their religion and sexual orientation. The question the paper is asking is whether extra information available online but not on the résumé leads to discrimination: Would applicants whose identity is not revealed in the application, but who appear to be Muslim (versus Christian) or gay (versus straight) on a popular social network, suffer unequal treatment?

To do so, they created distinct online profiles: one profile on a professional network site, and another profile on a social network site where the emphasis is on sharing photographs or leisure-related comments, not job opportunities. The profile on the professional network site was identical across treatments (even the photograph was the same). The name used by researchers (selected after careful testing) was one not commonly associated with a particular race or religion. That is, the name of the "Muslim candidate" was non-Arabic, but the candidate's religion could be inferred after some search on the social network site. Only the profile on the social network site contained clues (e.g., Christian versus Muslim or straight versus gay).

The experiment finds that only a small fraction of employers use social media to conduct additional inquiry about job candidates.[12] Given the limited search efforts by employers, the effects of group membership are generally small. The total effect of trait manipulation is not

[12]Measuring the exact number of visits to a social or professional networking profile is not possible for several reasons. However, using Google Adwords "Keyword Tool" statistics and Professional Network "Premiere" account statistics, the authors estimate that at most one-third of the employers tried to access the profile of the candidates.

statistically significant: 12.6 percent of applicants who appeared to be Christian received call-backs, compared to 10.9 percent of candidates who appeared to be Muslim. About 10.6 percent of candidates who appeared to be straight males received call-backs, and the share of call-backs for seemingly gay males was nearly identical.

The strength of this type of study is that researchers are able to study more naturally the impact of traits that traditionally are not revealed on a résumé. While some correspondence tests have tried to signal religious affiliation or sexual orientation through "extracurricular activities" described on CVs, this type of disclosure might reveal more than religion or sexual identity: the employers might be reacting to someone's activism regarding their religion or sexuality, not their religion or sexuality per se.

While Acquisti and Fong (2013) focus on the impact of gay status for males and religion, their methods could be used to study the effect of other interesting and until now mostly unexplored characteristics. For example, would the size of a candidate's network have an effect? Would employers infer that a "popular" candidate has valuable social skills? Would attractive-seeming candidates receive more call-backs, or would attempts to "choreograph" one's online presence be viewed as an undesirable trait? Would candidates who reveal their family status be treated differently than candidates who are more private? Clearly, online field experiments offer a rich landscape for studying "what employers want."

1.5 Limitations of Correspondence Studies

While correspondence studies address some key weaknesses of the audit methodology, they share other weaknesses with audit studies and have some unique limitations of their own

Both types of studies can only inform us about the average differences in hiring behavior. But we generally think that applicants care about the marginal response. Real job seekers are likely to adjust their behavior during the search process in a strategic manner: in other words, they will not apply for positions in a random fashion. So, while informative about discrimination on average in a given setting, correspondence and audit studies are not informative about discrimination at the margin, when real job seekers have fully optimized their job search strategy to the realities of the workforce. This is related to a criticism raised by Heckman and Siegelman in their contribution to *Clear and Convincing Evidence: Measurement of Discrimination in America* when they challenge the use of newspaper advertisements in audit studies, referring to previous

findings that most jobs are found through direct contact with a firm, or via informal channels like family and friends:

> [c]ollege students masqueraded as blue collar workers seeking entry level jobs. Apart from the ethical issues involved, this raises the potentially important problem that the Urban Institute actors may not experience what actually occurs in the these labor markets among real participants (Fix and Struyk, 1993).

Another drawback of field studies (both audit and correspondence) is that fictitious applicants typically only apply to entry-level jobs. There are a few exceptions, and some of the studies we describe above use applications to skilled and experienced positions. But the bottom line is that many jobs are never advertised, and the extent of discrimination in the workplace overall may be quite different from the discrimination that is measured at the entry point in the labor market.

Yet another limitation of field studies (both audit and correspondence) is that the outcome variables that can be studied are typically very coarse. In fact, in this regard, the correspondence studies are inferior to the audit studies. Most of the time, interview invitations or rental offers ("call-back rates") are the only outcomes captured by field experiments (one exception being Doleac and Stein (2013), who were able to track transactions – sales of iPods through local online markets – all the way to completion). Obviously, because there is no real applicant, the correspondence study methodology cannot be taken to the interview stage, job-offer stage, or wage-setting stage – or to the stage at which people sign a lease on an apartment. Theoretically, all of this can be measured when auditors are used. However, even audit studies do not allow one to track other important outcomes, such as work hours, working conditions, or promotions. The binary outcome in the typical correspondence studies (call-back or not) raises important issues about how to conduct some of the analysis. What should be inferred about discrimination for the employers that do not call back any of the fictitious applicants? Is that evidence of "symmetric treatment"? Riach and Rich (2002) argue that if both the majority and minority candidate are rejected, that does not constitute evidence of equal treatment. Only with more continuous outcome variables – ones that typically are not available to the researcher, such as the ranking of the job candidates by the employer – would it be possible to resolve this tension.

Both correspondence and audit studies have also raised ethical concerns. Employers' time is bound to be a scarce resource, and researchers who carry out these studies are using it without

the involved parties' consent. A positive take on this ethical issue is List (2009) who argues that "[w]hen the research makes participants better off, benefits society, and confers anonymity and just treatment to all subjects, the lack of informed consent seems defensible." However, many people outside the scientific community would probably disagree. (In fact, List (2009) refers to experiments where subjects are compensated; in the case of correspondence tests, we did not come across experiments where employers were actually compensated for their time.)[13]

Another underappreciated ethical issue is that when the "applicant" declines an offer, things other than the anticipated consumption of the employer's attention can occur. The employer may "learn" (become convinced) that applicants with the attributes similar to those of the fictitious candidate are unlikely to accept offers. If this really happens, it is possible that some real job applicants will be treated differently (possibly less favorably) due to prior communication with the researcher pretending to be a job candidate. But it also possible that after observing a rejection or two from fictitious candidates, an employer may end up having the impression that the market is tighter than he or she thought; screening could then become less intense, which might be beneficial for real jobless candidates (but potentially detrimental for the employers).[14]

A subtler criticism of the correspondence and audit methods by Heckman and Siegelman (1993) has been recently revisited by Neumark (2012). Heckman and Siegelman (1993) show that a troubling result emerges in audit or correspondence studies because the outcome of interest is not linear in productivity (as it might be for a wage offer), but instead is non-linear. That is, we think that in the hiring process firms evaluate a job applicant's productivity relative to a standard, and offer the applicant a job (or a call-back) if the standard is met. This non-linear relationship can raise issues for any inferences of discrimination based on call-backs if employers believe that Blacks and Whites differ in the variance of their unobserved productivity.

Consider for example the case where employers believe that the variance of unobserved productivity is higher for Whites than for Blacks. The correspondence and audit methods make

[13]The method of correspondence studies has also been taken to the dating market (e.g Ong and Wang (2015)). We do not review these contributions here because it is a bit difficult to talk about discrimination when referring to the choice of whom to date, but the ethical dilemma of putting fake applications on a dating website also seems particularly acute. As a conceptual aside, it is also not at all clear that one needs to create a fictitious profile on dating websites, as it is already possible for the researchers to observe exactly the same information that the user has when making a decision. There is thus no "unobserved" variable biasing the analysis and no information to be gained from fictitious résumés. The exercise can be performed with observational data (See Fisman et al. (2008), Hitsch et al. (2010), and Banerjee et al. (2009)). This makes the ethical concern particularly salient.

[14]This particular issue can be at least partially addressed by debriefing employers ex-post about them having been part of a research study.

black and white applicants equal on observable productivity characteristic X_1. However, no information is conveyed on a second, unobservable productivity-related characteristic, X_2. Because an employer will offer a job interview only if it perceives or expects the sum $\beta_1 X_1 + X_2$ to be sufficiently high, when X_1 is set at a low level the employer has to believe that X_2 is high (or likely to be high) in order to offer an interview. Even though the employer does not observe X_2, if the employer knows that the variance of X_2 is higher for Whites, the employer correctly concludes that Whites are more likely than Blacks to have a sufficiently high sum of $\beta_1 X_1 + X_2$, by virtue of the simple fact that fewer Blacks have very high values of X_2. Employers will therefore be less likely to offer jobs to Blacks than to Whites, even though the observed average of X_1 is the same for Blacks and Whites, as is the unobserved average of X_2. The opposite holds if X_1 is set at a high value: in this case, the employer only needs to avoid very low values of X_2, which will be more common for the higher-variance Whites. In other words, Heckman and Siegelman (1993) show that, even when there are equal group averages of *both* observed and unobserved variables, an audit or correspondence study can generate biased estimates, with spurious evidence of discrimination in either direction, or spurious evidence of its absence.

Building constructively on this criticism, Neumark (2012) shows that if a correspondence study includes variation in observable measures of applicants' quality that affect hiring outcomes, an unbiased estimate of discrimination can be recovered even when there are group differences in the variances of the unobservables. Neumark explains how his method can be easily implemented in any future correspondence study. All that is needed is for the résumés or applicants to include some variation in characteristics that affect the probability of being hired.[15]

Finally, it is remarkable that after literally dozens of correspondence studies, there has been only limited refinement of the methodology to help discriminate between different theories of the differential treatment that is being consistently observed. Employers must try their best to infer future productivity of a candidate based on limited information. That is, applicants who belong to different groups may experience different treatment even if discrimination as understood by Becker (differential treatment is motivated by prejudice) is absent and only statistical discrimination is at play. Attributes beyond those intended by the researcher may be inferred

[15]The method rests on three types of assumptions. First, it is based on an assumed binary threshold model of hiring that asks whether the perceived productivity of a worker exceeds a standard. Second, it imposes a parametric assumption about the distribution of unobservables. Lastly, it relies on an additional identifying assumption that some applicant characteristics affect the perceived productivity of workers, and hence hiring, and that the effects of these characteristics on perceived productivity does not vary with group membership.

by the recipient. For example, Fryer and Levitt (2004) suggest that black names may "provide a useful signal to employers about labor market productivity after controlling for information on the résumé." This is clearly true for age, as we noted, but this may also be true for race if the choice of a black name is a political statement by the parent, accompanied by a different attitude towards schooling and obedience. More broadly, as we already mentioned several times, even if in general employers do not see a particular identity as a sign of lower productivity (or want to discriminate based on it), they may infer something from the fact that the person is wearing it on their sleeves. After all, there was no difference in call-back rate according to either religion or sexuality when the information was available to the employer online, but not directly reported in the résumé (Acquisti and Fong, 2013).

The only approach that has been used repeatedly by researchers to try to separate statistical from taste-based discrimination has been to compare differential gaps in outcomes between pairs of minority and non-minority applicants with weaker or stronger productivity attributes on their résumés or applications. As more productivity-relevant information is included on the résumé, average differences in unobservable characteristics between the minority and non-minority applicants are reduced, and statistical discrimination should also be reduced; however, it is clear that this remains a very indirect way to try to isolate taste-based discrimination among employers or landlords.

In this regard, a recent paper that breaks the mold of the typical correspondence study and deserves particular attention is Bartoš et al. (2013). This paper is remarkable in its ability to push the correspondence study methodology forward, think beyond the pure call-back data, and refine our theories of discrimination.

The paper links two important ideas: attention is a scarce resource, and lack of information about individual candidates drives discrimination in selection decisions – or, in other words, statistical discrimination is an important factor in selection decisions. While the existing models of statistical discrimination implicitly assume that individuals are fully attentive to available information, the paper develops and tests a model in which knowledge of minority status impacts the level of attention to information about an individual and how the resulting asymmetry in acquired information across groups – denoted "attention discrimination" – can lead to discrimination. In particular, the authors argue that in markets where only a small share of applicants is considered above the bar for selection, such as the labor market, negative stereotypes are

predicted to lower attention. On the other hand, the effect is opposite in markets where most applicants are selected, such as the rental housing market.

Bartoš et al. (2013) test for such "attention discrimination" in two field experiments: one in the labor market and one in the rental market, where they can monitor the decision-maker's information acquisition about applicants through visits to hyperlinks containing résumés and personal websites (respectively). They created personal websites for fictitious applicants and submitted rental applications in the Czech Republic, and job applications in both Germany and the Czech Republic. The advantage of using hyperlinks to résumés and personal sites is that the researchers were able to track the *exact number of visitors* to the personal profile, and therefore the share of landlords and employers who allocated additional attention to an applicant. Hence, the study was able to assess whether a minority-sounding name 1) leads to differential call-backs and 2) causes less or more search.

Like the prior literature, Bartoš et al. (2013) find evidence of discrimination against minority applicants in both the housing and labor markets. Most interesting, though, are their findings regarding attention allocation. In the labor markets in both Germany and the Czech Republic, employers put more effort in opening and reading résumés of majority compared to minority candidates. In contrast, in the rental housing market, landlords acquire more information about minority compared to majority candidates through their personal sites.

The findings can best be explained by a model where attention is endogenously determined by the type of the market. When the choosing entity needs to select "top candidates," it will allocate attention to candidates belonging to the group that, according to its priors, is stronger. In markets where most candidates are accepted, some kind of a threshold rule might be used, and the choosing entity will want to eliminate the weakest candidates. In that case (e.g. a housing market), more attention would optimally be allocated to members of the group that is a priori viewed less favorably. The model implies persistence of discrimination in selection decisions, even if information about individuals is available and there are no differences in preferences. The model also implies lower returns to employment qualifications for negatively stereotyped groups (as their credentials are less likely to be reviewed). From a policy perspective, the model and results of this paper also highlight the crucial role of the timing of when a group attribute is revealed.

Bartoš et al. (2013) represents a great example of how the résumé study infrastructure can

be pushed forward to deliver deeper learning, cleaner links to specific theories of why differential treatment is taking place, and suggestions about policies that might be most effective to address it. More efforts along these lines would help revitalize this literature.

We now turn to other approaches to measuring discrimination, often more "lab-based" and more closely tied to a particular model of the root of discrimination.

1.6 Implicit Association Tests

The Implicit Association Test (IAT) is a computer-based test that was first introduced by Greenwald, McGhee, and Schwartz (1998). Developed by social psychologists Greenwald, Nosek and Banaji, as well as other collaborators, the IAT provides a method to indirectly measure the strength of association between two concepts. This test relies on the idea that the easier a mental task is, the quicker it can be performed. When completing an IAT, a subject is asked to classify, as rapidly as possible, concepts or objects into one of four categories with only two responses (left or right). The logic of the IAT is that it will be easier to perform the task when objects that should get the same answer (left or right) somehow "go together".[16]

The typical IAT consists of seven "phases," including practice phases to acquaint the subject with the stimuli materials and rules. Consider for example an IAT designed to assess association strengths between categories of black and white and attributes of good and bad. The practice phases are used with the materials and sorting rules. In the first, subjects would only be presented with faces as stimuli and be asked to assign white faces to one side and black faces to the other; in the second, subjects would only be presented with words as stimuli and be asked to assign pleasant words to one side and unpleasant words to the other. In the test phases, subjects are asked to simultaneously sort through stimuli representing the four concepts (black, white, good, bad) but with again only two responses (left side or right side). In two of the test phases (the "stereotypical" test phases), items representing white and good (e.g., white faces and words such as wonderful) need to be placed on one side of the screen, and items representing the concepts black and bad (e.g., black faces and words such as horrible) on the other. In the other two test phases (the "non-stereotypical" phases), items representing the concepts of black and good need be placed on one side of the screen, and items representing the concepts of white and bad on the other. The extent to which an individual dislikes black faces (in this case) is then

[16] See Lane et al. (2007) for an excellent introduction to IATs.

measured by the difference in milliseconds in response time between the stereotypical phases and the non-stereotypical phases.[17]

Two broad kinds of IAT are pertinent to discrimination: if attitudes or overall preferences are the issue, the category (e.g. black/white) is associated with words that represent good/bad (as in the example we just gave). Alternatively, one may be interested in the association between a category (e.g. male/female) and a particular trait or attribute (e.g. career/family).[18] The first kind is called an attitude IAT, and the second a stereotype or belief IAT. Other types include self-esteem IATs (e.g. categories are self and other and words are either positive or negative).

Since the publication of the original IAT, there have been hundreds of IAT studies, many of which try to capture attitudes that could give rise to discrimination (against black people, Muslims, women etc.), or phenomena more akin to statistical discrimination (women and math, women and career, women and politics, etc.). The IAT has been extremely influential both within and outside academic psychology. Greenwald, McGhee, and Schwartz's original 1998 article introducing the IAT has 6689 citations in Google Scholar, as of August 2015. The findings of IAT research on discrimination have been cited to propose changing the law, educating judges and students, etc. IATs are used increasingly as a convenient tool to measure whether attitudes respond to any particular intervention, since they can be conducted remotely, with large samples of online participants to experiments. As such, they are often used as endpoints in psychology experiments, as experimentation moves to online platforms.[19]

There are a number of meta-analyses, review articles, and critical papers on the use of IATs. It is not in the scope of this paper to review all of this literature, however the key question that is raised is what the IAT actually picks up and, relatedly, whether it is effectively associated with other predictors of discriminatory behavior, and discriminatory behavior itself. Some individual studies show promising links. For example, implicit bias predicts a more negative judgment of ambiguous actions by a black target (Rudman and Lee, 2002), as well as more negative non-verbal "microbehavior" (less speaking time, less smiling, etc.) during an interaction with a black subject (McConnell and Leibold, 2001). This is important, as these microbehaviors are often posited to be the channel through which implicit bias would translate into different behavior,

[17] In practice, of course, a number of choices must be made about how to use the data, and this is reviewed in Greenwald, Banaji, and Nosek (2003)

[18] For example, Nosek, Banaji, and Greenwald (2002) find stereotypical associations connecting male terms with traits related to science and a career, whereas female terms are found to be associated with liberal arts and family.

[19] For examples of IATs used as endpoints, see Lai et al. (2014)

even among people who do not report explicit discrimination.

Some studies have also shown some mechanisms for those effects, e.g. showing that participants who exhibited greater implicit distaste of black people were more likely to detect aggression in a black (but not white) face (Hugenberg and Bodenhausen, 2004). Only a few studies have investigated whether these differences in implicit attitudes are associated with different behaviors in the field. In Atlanta and Boston, doctors with stronger anti-black implicit attitudes were less likely to prescribe thrombolysis for myocardial infarction to African-American patients, compared to white patients (Green et al., 2007). Rooth (2010) tried to relate the behavior of recruiters in a correspondence study in Sweden (focusing on Arab-Muslim versus Christian) to recruiter-level measures of implicit discrimination they collected later. Unfortunately they were only able to interview 26 percent of the recruiters they were targeting, but among those, they did find a correlation between implicit distaste of Arabs as measured in an IAT test and the tendency to not call back a resume with an Arab-Muslim name on it.

An initial meta-analysis, conducted on 122 research reports, found that the IAT does seem to be capturing something about attitudes, perhaps more accurately than self-reports (Greenwald et al., 2009). They show that there is a strong correlation between implicit and more standard explicit measures. Moreover, the IAT appears to be a better predictor of actual behavior than explicit reports, particularly for sensitive subjects such as racial preferences (for which they have 32 samples with IAT measure, explicit measure, and questions about behavior).

However, a more recent meta-analysis by Oswald et al. (2013) questions these initial findings. Using a larger sample (which includes newer studies as well as some studies that were omitted from the earlier meta-analysis), and a slightly different aggregation method, they find much lower correlation of the IAT with various measures of discrimination than had been initially found in the 2009 meta-analysis. Explicit measures perform equally poorly, to be sure, but not much worse.

Beyond this debate (which is probably the core one to be had), IATs have been subject to a number of criticisms and questions, mainly regarding their interpretation. First, to the extent that they differ from explicit attitudes, do they reflect something "deeper" about the individuals and are they more "true" than the self-description in any sense, or simply another type of attitude? Interestingly, the meta-analysis by Oswald et al. (2013) shows that there is in fact a strong correlation between different brain activities when seeing black and white faces

and the IAT. This suggests that the IAT does reflect something fundamental about psychological processes. But our behavior is mediated by the social environments, exactly as our answer to a question on prejudice is mediated by this environment.

So do IATs really identify prejudice or just some raw psychological "material" that we then transform? What does it mean for someone to feel that there are not prejudiced against Blacks but have their IAT showing automatic white preferences (Arkes and Tetlock, 2004)? On this last question, Banaji, Nosek, and Greenwald (2004) argue that conscious unbiased attitudes cannot be relied upon in all circumstances, and that IATs may capture unconscious attitudes that may be more relevant in explaining behaviors in other circumstances. Hence they reject the idea that "if prejudice is not explicitly spoken, it cannot reflect a prejudicial feeling" (Banaji, Nosek, and Greenwald, 2004). In this respect, though, the low correlation between the IAT and micro-behavior is a bit troubling, as this theory would suggest that the unconscious bias translates into actual acts of discrimination via unconscious behavior.

Also, do IATs measure the prevalent culture or individual attitudes? For example: if a person identifies women with family more than with career, is she making a value judgment or stating, in a sense, a fact of life?

Whatever the resolution of these debates, the measured implicit attitudes vary considerably across people, and the robust correlations – between implicit and explicit attitudes, between the different IATs in similar domains – do seem to indicate IATs capture some signal about the individual. This does not mean that the IAT can be considered a reliable measure of the attitude of any particular individual (at best, it measures attitudes with considerable noise). However, it does mean that the IAT may be a good measurement tool for the propensity for groups to discriminate towards each other.

In this context, whatever the predictive value of the IAT for behavior, the extent to which it is affected by a particular manipulation is of interest. As economists, we may be more interested in the extent to which attitudes can be influenced (by experiences, the environment, or specific interventions), than in their pure measurement at a point in time. Using IATs as an outcome variable also helps side-stepping the question of whether they represent any deep truth about anybody: while the signal may be noisy, to the extent that there is signal, this may be a useful measurement tool. As noted, after more than a decade of using the IAT mainly as a descriptive tool, studies in psychology started using them as outcomes. For example, Lai et al. (2014) set

up a research contest on de-biasing, where teams are given a budget of five minutes to interact with participants, and the outcome is the scores on a black-white attitude IAT.

In recent years, economists have also started using IATs as dependent variables. For example, Beaman et al. (2009) design and implement two IATs in West Bengal, India, to measure preferences towards female leaders, and stereotypical association of women with domestic rather than political activities. They then examine the impact of exposure to female leaders on these two measures (we will discuss the results below in Section 1.7).

Lane et al. (2007) provide detailed and helpful instructions on how to build an IAT. The software that is needed to construct and analyze the test (millisecond software) is available for purchase. IATs can be designed with only verbal or image stimuli for subjects who are not literate (this is what Beaman et al. (2009) use) and although they are more difficult in populations that have had no experience with computers and for older participants, they can be a very useful tool. As studies increasingly use electronic data collection methods (on tablets or notebooks), the extra cost of adding an IAT diminishes.

Of course, the debate in psychology does not suggest that the IAT should be considered a "magic bullet," suitable to replace any other measure of discrimination. In particular, it is probably not a substitute for good measures of actual behavior in policy interventions. Nevertheless, it can be an extremely useful intermediate variable, to understand the mechanisms beyond a result (in Beaman et al. (2009) the final endpoint of interest is actual voting), or potentially, if collected beforehand, as a covariate of interest. For example, in Glover, Pallais, and Pariente (2015) the IAT is used as a proxy measure of latent employer discrimination (see further details in Section 2.2.2).

1.7 Goldberg Paradigm Experiments

Goldberg Paradigm experiments are laboratory versions of audit or correspondence studies. They are named after a 1968 experiment by Goldberg where students graded written essays, which were identical except for the male or female name of their author (Goldberg, 1968). This initial experiment demonstrated a bias: female got lower grades unless the essay was on a feminine topic. Since then, a large literature in psychology has used the Goldberg Paradigm to identify

discrimination against different groups, and in particular in the resistance to female leaders.[20]

In the typical lab experiment, a group of subjects is asked to review a vignette, describing the behavior of a female or male manager (for example), or witness a confederate (male or female) simulating a leadership situation. The participants are then asked to evaluate the leader's competence, or to say whether they would have liked to have them as leader for a task they may collectively perform. Reviewing a large number of such studies, Eagly, Makhijani, and Klonsky (1992) do not find that, on average, female leaders are evaluated significantly more negatively than male leaders. However, there are in some circumstances where they do find that female leaders are evaluated more negatively: for example, when the leadership was carried out in a masculine style (in particular when the leader was projected to be authoritative). This supports Eagly's hypothesis of "role congruence": what people dislike is when women behave in a non-feminine way. Since strong leaders must be assertive, but women must be demure, it makes it difficult for women to be appreciated as strong leaders.

The fact that the circumstances are artificial, and answers have minimal stake associated with them, make those experiments less relevant, on their own, than field-based correspondence tests. But one advantage of the Goldberg-style experiments is that they can be easily, and finely, manipulated, which makes them good outcome measures in field research (or field experiments). They can also be easily added to a standard survey instrument. For example, Beaman et al. (2009) seek to find out how discrimination against female leaders is affected by prior exposure. They administer two Goldberg-style experiments. In one, they ask the participants to listen to a speech by a political leader, which is read either by a female or a male actor (note that it is important that there are several male and female actors). In the second one, they discuss vignette where women or men leader make decisions that are either pro-male (investment in irrigation) or pro-female (investment in drinking water). Each individual receives a randomly selected version of the speech and vignette. The randomization is stratified by village, and hence by prior exposure to a female leader (due to a policy of gender reservation). While this does not tell us the extent to which any single person discriminates, one can learn whether, on average, exposure to a female leader affects the extent to which individuals give lower grades to women in response to the same speech or vignette. Beaman et al. (2009) find that both men and women,

[20]See Eagly, Makhijani, and Klonsky (1992) for a review and meta-analysis of the literature on such resistance to female leaders.

but men more than women, tend to discriminate against female leaders (additional results from this study are discussed further in Section 3.1.2).

1.8 List Randomization

Like correspondence tests or Goldberg experiments, list randomization (also known as item count technique, unmatched count, or list response) do not provide a measure of individual bias, but can provide an estimate of the extent of discrimination in a population. They are a way of eliciting accurate answers to questions of discrimination in the presence of social desirability bias. The idea is to present the subjects with a list of N statements which are generally non-controversial, but could be true or false (e.g. I had coffee at breakfast; I like popcorn).[21] Then, a randomly selected group of people is asked a potentially controversial statement (e.g. "I would be upset if an African-American family moved next door") on top of the N non-controversial statements. The subject only states the number of statements with which he or she agrees. Comparing the fraction of yes among those who got N and those who got N+1 statements gives a good measure of discrimination. And clearly no one (including the interviewer) will know how a given subject answered the controversial statement. Unlike the IAT, this method will not reveal biases that are unconscious or biases that the subject wants to deny even to himself or herself, but it will prevent the results from being affected by social desirability bias.

Early applications of list randomization to measure discrimination are Kuklinski, Cobb, and Gilens (1997a) and Kuklinski et al. (1997b). Both studies found considerable racial prejudice in the American South using list randomization techniques (though not in the North). Furthermore, they found higher level of measured discrimination using this method than using direct elicitation methods; for example, respondents are more likely to disagree with a statement such as, "I am comfortable with a black family moving next door" when asked via list randomization than with a traditional survey. Likewise, Coffman, Coffman, and Ericson (2013) show that stated discrimination against gay populations is much lower in response to a direct question in the control group than when it is elicited through the list randomization method. For example, respondents were 67 percent more likely to express disapproval of an openly gay manager at work when the question is part of a list than when the question is asked directly.

[21] As we explain below there is a tension in the choice of those questions: for maximum precision they should be behavior that almost everyone says yes or no to, but then they do not give any cover to the subject.

A few papers have used the method to elicit attitudes towards presidential candidates. Kane, Craig, and Wald (2004) find no discrimination against a Jewish presidential candidate (Joe Lieberman). Martinez and Craig (2010) find that few Whites in Florida seemed distressed by the possibility of having a black president. However, list randomization revealed much more opposition towards the idea of a female president than opinion polls (Streb et al., 2008). As noted above, several studies suggest that the randomized list technique yields different answers than direct elicitation. In a meta-analysis across 48 comparisons of direct report and list randomization, Holbrook and Krosnick (2010) found that 63 percent of the estimates for socially undesirable behaviors were significantly larger when elicited through list randomization. On the other hand, responses on non-sensitive behavior tend to be more similar (Tsuchiya, Hirai, and Ono, 2007).

The list randomization method is however not without issues. As we alluded to earlier, there is a fundamental tension between precision (which would require having statements to which everybody responds yes or no) and providing "cover" to the subject (which would require the opposite). The implication is that the results from list randomization methods are often quite imprecise. Gosen (2014) has also shown results tend to systematically depend on how many noncontroversial statements are included in the list, although the opposite was found in Tsuchiya, Hirai, and Ono (2007).

In summary, list randomization could be a promising method to measure discrimination as it less subject to social desirability bias, but since few economists have used it,[22] more work needs to be done to ascertain its usefulness in the field. In comparison to other indirect methods, list randomization is often more simple to administer (both for surveyors and respondents) but risks having low power (Droitcour et al., 1991). It would be interesting to see more research comparing measures of discrimination obtained through list randomization compared to an IAT or Goldberg style experiment. It would also be interesting to compare how noisy these different measures are. The fact that the list randomization method can only provide an aggregate (and not individual) measure of discrimination complicates its use as an outcome variable (say, for a randomized experiment), but no more than any of the other methods we have already discussed that also only give group-level outcomes, such as the Goldberg-style experiments.

[22] See Karlan and Zinman (2012) for an example of an application in economics.

1.9 Willingness to Pay

A key prediction of Becker's model of taste-based discrimination is that people should be willing to pay to interact with people of their own group. Somewhat surprisingly this prediction has not given rise to a large literature trying to evaluate the willingness to pay to discriminate. As we noted, the correspondence and audit tests tend to be based on a binary measure (interview or not, hire or not).

Until recently, the body of work that came closest to measuring such willingness to pay was a literature on the "beauty premium" motivated by Hamermesh and Biddle's (1994) finding that workers with better-than-average looks earn 10 to 15 percent higher wages. Analogous to the black-wage race gap, the beauty premium could be due to the fact that more beautiful workers are more productive, say because consumers prefer to interact with beautiful people (Biddle and Hamermesh, 1998; Pfann et al., 2000), or because beautiful people are more confident. Or employers may be wrong in their belief.

Mobius and Rosenblat (2006) set up a laboratory experiment where undergraduates and graduates from Tucuman, Argentina were randomly assigned into groups of "employers" and "workers." In the experiment, "employers" had to hire "workers" to perform a maze-solving task. After a practice test (which was recorded and became the digital "resume" of the worker) and a question where the workers estimated the number of mazes they could solve in 15 minutes, each worker was matched to five employers, who saw either (1) just the resume, (2) the resume and a photo, (3) the resume and a phone interview, or (4) the resume plus an interview, plus the photograph.[23] The employers in turn saw five workers, and for each of them decided how many mazes they thought the worker could solve. This estimate contributed to the employer's own payment. It also entered into the calculation of the actual wage of most of their workers.

Mobius and Rosenblat show that productivity at the task is not affected by beauty (as evaluated by 50 high school students on the basis of the photograph), although worker confidence is. A rise of one standard deviation in beauty increases confidence by 13 to 16 percent. However, employers are willing to pay more employees who are considered to be more beautiful: in all the treatments where they can see beauty, employers are willing to pay workers more. The premium ranges between 12 and 17 percent depending on the treatment. Decomposing the

[23]The wages of "workers" were affected by the difference between their estimation of the number of mazes and what they actually competed, therefore they were incentivized to tell the truth.

beauty premium by comparing treatments, the authors estimate that 15 percent is due to the confidence channel and 40 percent each through the visual and oral stereotype channels (the fact that beauty still affects wages when the employer does not see the employee but talks to her on the phone indicate that beauty is correlated with speaking skill, perhaps another feature of the beauty channel). Interestingly, employer's estimated productivity is not affected by whether or not they know that it will actually contribute to the worker's wage. This suggests that there is little pure taste-based discrimination in this lab experiment. Employers give a premium to beautiful people because they believe (wrongly) that they will be more productive.

There are a number of limitations of this experiment, not least of all, from the point of view of this chapter, that it is a lab experiment. It is also limited to a one-shot interaction at the hiring stage. Nevertheless, it sets an interesting template for what a field experiment leveraging this methodology might look like, and in particular does an excellent job laying out the various pieces that are needed to establish discrimination and understand the mechanism behind it.

One paper which has recently followed in Mobius and Rosenblat's footstep is Rao (2013), which seeks to measure the extent to which well-off kids in India discriminate against poorer kids (in order, as we will discuss in more detail in Section 3.2, to estimate the extent to which any such discrimination is affected by forced exposure to poorer kids through an affirmative action program in education). To do so, Rao sets up an ingenious field experiment, based around team selection for a relay race. First, students from a rich private school and a poor public school, who were all present at a sporting event to support their classmates, were randomized in different sessions with different prizes for winning the race (from Rs50 to Rs500, which are very high stakes). After mixing for 15 minutes, they watched a series of one-on-one sprints (most of them pitting a poor student against a rich one) and then each was asked to indicate on a worksheet which of the two he wanted as teammate for the relay race. After these choices were revealed, one of the choices was picked, the teams were formed, and the relay race was run. To make sure that there was a "cost" to picking out a poorer students (if students did not like them), students had to spend two hours socializing with their teammate, which was announced prior to team selection. This experiment has a number of clever features. It presents children with a real choice, and by varying the stakes it makes it clear how much (on average) students are willing to sacrifice to avoid interacting with a poor student. The sprint phase entirely and unambiguously reveals ability, so the set-up is targeted to pick up pure taste-based discrimination (e.g. dislike

of hanging out with a poor teammate).

The results show that there is substantial taste-based discrimination in this context: in 19 percent of the cases where the poor students is the fastest, rich students prefer to pick the rich student as a teammate anyway.[24] Discrimination does decline as the stakes increase: discrimination falls from 35 percent with the lowest stake, to 27 percent with the intermediate stake, and 5 percent in the highest stake. Fitting a structural model to the data, Rao estimates that, for students without prior exposure to poor classmates, the distaste of interacting with a poor student is worth Rs37. That is, a student is willing to give up to Rs37 in expected prize money to hang out with a rich student rather than a poor one.

2 Consequences of Discrimination

2.1 Self-Expectancy Effects

2.1.1 Stereotype Threat and Underperformance

Models of statistical discrimination explain the differential treatment of disadvantaged groups in say, hiring decisions, due to employers' inability to perfectly predict a given worker's future productivity and hence their rational decision to assign some weight to the average productivity of the worker's racial group. For example, African-Americans as a whole are categorized as less productive than Whites, and employers take this supposed average difference in productivity into account when deciding whether or not to hire any African-American job candidates, given their inability to precisely predict each specific candidate's future productivity.

While discrimination emerges under the logic above as a consequence of average differences in productivity across groups, research in social psychology has provided convincing evidence for the reverse causal channel. In particular, the simple process of categorizing or "stereotyping" some groups as less productive appears to cause these groups to be less productive. This research suggests that individuals from some groups may suffer negative performance outcomes (such as lower test scores or less engagement with academics) because of the burden of the "stereotype," or "stereotype threat" (Steele and Aronson, 1995).[25] The key conjecture is that the threat of being

[24] In contrast, if a rich student is the fastest he is picked 97 percent of the time, and among two students of the same background, the fastest is picked 98 percent of the time.

[25] Defined by Steele and Aronson (1995) as a "risk of confirming, as self-characteristic, a negative stereotype about one's group."

viewed through the lens of a negative stereotype can create an anxiety that disrupts cognitive performance.

In a seminal study, Steele and Aronson (1995) demonstrated in a lab setting that inducing stereotype threat – by asking test takers to indicate their race before the test – significantly undermines African-Americans' performance on intellectual tasks. They also showed that reducing stereotype threat – by convincing test takers that the test was not being used to measure their abilities – can significantly improve African-Americans' performance, dramatically reducing the racial gap.

Numerous lab studies have since replicated the effects of stereotype threat both with respect to social identities other than race (e.g., gender, income class, etc.) and with respect to mediating outcomes (such as blood pressure, heart-rate variability, performance expectations, effort, etc.).[26]

The rest of the social psychology research on the stereotype threat has also focused on documenting methods that can undo or undermine the threat of the stereotype. One line of research has addressed the underlying message of the stereotype – that stereotyped individuals are inherently limited because of their group membership. Thus, if participants can be convinced that intelligence is not a fixed trait, but a malleable quality that can be increased with hard work and effort, they may be less prone to stereotyping. Levy, Stroessner, and Dweck (1998) present evidence consistent with this idea. Descriptively, they find that people holding an entity theory of intelligence (i.e. intelligence is a fixed trait) made more stereotypical trait judgments of ethnic and occupational groups than those who believed that intelligence is malleable. Moreover, in a small lab experiment, they found that manipulating implicit theories affected level of stereotyping, at least temporarily. In the experiment, 155 introductory to psychology students were randomly assigned to a fake "scientific" article that either presented evidence for an entity (fixed) or an incremental (malleable) view of personality. After reading the article, they were presented

[26]While most of these lab studies have been conducted by social psychologists, a few have been performed by economists. For example, Dee (2009) ran a lab experiment with student-athletes and non-athlete students at Swarthmore College, randomly assigning some of them to a treatment that primed their awareness of a stereotyped identity (i.e., student-athlete). He finds that the treatment reduced the test-score performance of athletes relative to non-athletes by 14 percent. Also, Hoff and Pandey (2006) present evidence from a caste priming experiment in Uttar Pradesh, India. Among 321 high-caste and 321 low-caste junior high school male student volunteers, there were no caste differences in performance in an incentivized maze-solving task when caste was not publicly revealed, but making caste salient created a large and robust caste gap in performance. However, the mechanisms for the underperformance in this case seems quite different from the hypothesized mechanism in the social psychology literature. In particular, the authors find that when a nonhuman factor influencing rewards received for the maze-solving task (a random draw) was introduced, the caste gap disappeared. The results suggest that when caste identity is salient, low-caste subjects anticipate that their effort will be poorly rewarded.

with questions in which they had to rate the extent to which a series of 15 traits accurately describe certain occupational groups (teachers, doctors, lawyers, and politicians) and ethnic groups (African-Americans, Asians, and Latinos). To try to reduce the likelihood of participants recognizing the link between the article and the questions, the researchers told participants that the questions were for a separate study, and that they would be asked questions on the content of the article later. The experiment found a small but significant effect: those who read the article that argued for fixed personality were less likely to believe traits can change, and more likely to rate stereotypical traits as highly descriptive of their respective groups.

Similarly, in a lab experiment, Aronson, Fried, and Good (2002) assign white and black students to one of three conditions to assess the impact of an intervention designed to reduce stereotype threat. In two conditions, students were asked to write a letter of encouragement to a younger student who was experiencing academic struggles. In one of these conditions, students were prompted to endorse a view of intelligence as malleable, "like a muscle" that can grow with work and effort. In the second condition, students endorsed the view that there exist different types of intelligence. The third condition served as a control condition in which students were not asked to compose a letter. Several days after the intervention, all students were asked to indicate their identification with and enjoyment of academics. Results showed that black students in particular were more likely to report enjoying and valuing education if they had written a letter endorsing malleable intelligence. In addition, grades collected nine weeks following the intervention were significantly higher for Blacks in the malleable intelligence condition. Whites showed a similar, though statistically marginal, effect.[27]

While most research on stereotype threat, and how to undo it, has taken place in the lab, a few interesting field studies have also been conducted by social psychologists. Schools, where test-taking and performance measurements are part of normal operations, have provided a natural setting for much of this field research.

Good, Aronson, and Inzlicht (2003) performed a field experiment to test methods for helping female, minority, and low-income adolescents overcome the effects of stereotype threat and, consequently, improve their standardized test scores. Specifically, seventh-grade students in

[27]It is important to note that for this experiment, while the randomized interventions took place in the lab, outcomes are measured 1) on naturally occurring tasks outside the lab and 2) quite a long time after the interventions took place; both of these features are important strengths of this experiment compared to the standard "stereotype threat" lab experiments.

the experimental conditions were mentored by college students who encouraged them either to view intelligence as malleable or to attribute academic difficulties in the seventh grade to the novelty of the educational setting. Results showed that females in both experimental conditions earned significantly higher math standardized test scores than females in the control condition. Similarly, the students – who were largely black or Hispanic and low-income adolescents – in the experimental conditions earned significantly higher reading standardized test scores than students in the control condition. Blackwell, Trzesniewski, and Dweck (2007) based a field experiment on the laboratory finding on "malleable intelligence." They randomly selected half of a group of 95 mainly African-American and Hispanic seventh graders to participate in an eight-week training on the theory of malleable intelligence based on interventions that had been successful in the lab (25 minutes per week, in the students' classroom). In the control condition, student also received small group coaching, but not on this theory. Students in the experimental conditions obtained higher grades.

Good, Aronson, and Harder (2008) also conducted a field experiment where they explored stereotype threat and its negation in high-level college math courses that typically serve as gateways to careers in math and science. Male and female students in the last course of an advanced university calculus sequence were given a practice test containing items similar to those found on the Graduate Record Examination (GRE) standardized test. All students were told that the test was "aimed at measuring your mathematical abilities" (stereotype threat) but half of the students additionally were assured that "this mathematics test has not shown any gender differences in performance or mathematics ability" (stereotype threat negation). Test performance was higher for women than men in the stereotype threat negation condition but was equivalent in the stereotype threat condition.

In a related field study, Cohen et al. (2006) reduced the black-white GPA gap among low-income middle school students by affirming the students' self-concepts (and presumably inoculating them from stereotype threat) at the beginning of the school term. The intervention is very light touch: students are asked to write a series of short essays focusing on a self-affirming value. The authors first found short-term impacts, and then, most remarkably, fairly large long-term impacts: over two years after the intervention, the average GPA of an African-American student who participated in the essay writing was 0.24 higher than that of the control group (Cohen et al., 2009). Most of the effects is concentrated among those who were initially low achieving.

These are remarkable numbers, especially since the effects of most education interventions tend to fade. The authors speculate that the long-term benefits may be due to the fact that initial psychological state sets out a self-fulfilling trajectory.

Cohen's work has since then been replicated and extended in other similar contexts, and Dweck and Cohen's initial insights have helped jumpstart a subfield of psychology called "mindset studies." Yeager et al. (2014) experiment with "wise feedback," an intervention where high school students are given critical feedback on their written work, alongside with a note emphasizing the high standard of the teachers and the belief that the student can succeed. The intervention reduces mistrust among African-American students and improves the quality of the final product.[28]

2.1.2 Identity and Preferences

The "stereotype threat" literature can be viewed as part of a broader literature on how self-identity considerations may affect behavior and preferences of disadvantaged groups and ultimately may perpetuate gaps in economic outcomes. The same way that women may do poorly on a math test when reminded of their gender (due to the anxiety-inducing burden of the stereotype of "girls not being good at math"), they may also show low risk preferences when reminded of their gender (if nurtured with the behavioral norm that "girls should not take risk" by gender-biased parents and/or teachers).

Against the backdrop of a large literature in social psychology that has tested the self-categorization theory and the cognitive mechanisms through which it operates,[29] a few recent papers in economics have leveraged the lab environment to learn more about how various social identities relate to preference parameters, such as risk, time and social preferences.

For example, Benjamin, Choi, and Strickland (2010) explore the effect of racial and gender category norms on time and risk preferences. In a laboratory setting, they study how making salient a specific aspect of one's social identity affects subjects' likelihood to make riskier choices, or more patient choices. From a methodological perspective, the study consists of temporarily making more salient ("priming") a certain social category (as is done in the "stereotype threat"

[28] A background paper written for a White House conference gives a good overview of the literature on mindset studies (Yeager et al., 2013).

[29] See for example Reicher and Levine (1994), Forehand, Deshpandé, and Reed II (2002), and LeBoeuf, Shafir, and Bayuk (2010)

literature) and seeing how the subjects' choices are affected. For example, the gender identity salience manipulation is done through a questionnaire included in the beginning of the experiment in which subjects are asked to identify their gender and their opinion regarding living on a coed versus single-sex dormitory floor. The study uncovers some interesting patterns with respect to racial identity. For example, priming a subject's Asian-American identity makes the subject more patient. Hence, an Asian-American identity might partly contribute to the higher average level of human capital accumulation in this racial group.

However, making gender salient appears to have no significant effects on either men's or women's patience, or their level of risk aversion. Of course, it is possible that the priming performed in this experiment was too weak to temporarily affect preferences. In other words, it is difficult to affirmatively conclude from these non-results that gender identity norms are not culturally reinforcing whatever biological differences may exist between the sex in the willingness to take risks.

Another lab study aimed at assessing how social preferences are affected by gender identity is Boschini, Muren, and Persson (2009). The question under study here is whether gender identity priming affects subjects' level of altruism. The experiment consists of comparing behavior in a dictator game for subjects whose gender identity has been primed versus not primed. The results indicate that the priming does affect behavior but only when the subjects are assigned to mixed-gender groups. Moreover, the effect is driven by males: men are sensitive to priming and become less generous in a mixed-gender setting when primed with their male identity. Women do not appear to respond to the treatment.

As far as we are aware of, no field experiment exists on how social identity affects preferences and behaviors outside of the mindset literature discussed above, which focuses on education and on adolescents. It seems worthwhile for future research to consider such work. Interventions might be designed to emphasize a "default" social identity that may be counterproductive for that social group's performance against an "alternative" social identity. For example, while deciding to work hard towards completing college coursework for a young black father might be uncool because it is "acting too white," the decision might resonate much more when his identity as a "father" is being primed. Moreover, specific interventions might be designed to simply undo or undermine the power of the social identity norms when they work towards reinforcing differences in behaviors and outcomes between groups. If women decide against applying for

a job in a high-risk but also high-return occupation because of internalized conservative social norms about "what is appropriate work for a woman," it might be possible to undermine the pull of this conservative norm with counteractive "messaging," in the same spirit as what has been done to undermine the burden of the stereotype in the "stereotype threat" literature. Such interventions might be particularly powerful if the timing of the counteractive "messaging" is close to when women are making these important career choices (e.g. when applying for school or on a job search website, or when considering which contact in their LinkedIn network to reach out to).

2.2 Expectancy Effects and Self-Fulfilling Prophecies

2.2.1 Pygmalion and Golem Effects

Suppose minority and majority workers have similar inherent abilities. How could differential beliefs about their abilities persist? One explanation is that employers' beliefs that minorities are on average less productive are self-reinforcing (Arrow, 1973; Coate and Loury, 1993). This could happen for two reasons. First, minority and majority workers may rationally make different skill investment or effort choices in the face of the beliefs of their employers. A minority worker may see less value in investing in her skills if she knows that the employers will be slow in updating their beliefs, and hence less likely to promote her. Second, the employers themselves may invest less in the minority workers (e.g. investing in training) if they do not believe that the workers will be "up to the task." In both cases, employers' beliefs about minority workers will be self-fulfilling.

The social psychology literature offers multiple demonstrations of such self-fulfilling prophecies.[30] Interesting, most of these demonstrations took place in the field.

The earlier work on self-fulfilling prophecies in the social psychology focused on how heightened expectations can be self-fulfilling. In a seminal study, Rosenthal and Jacobson (1968) conducted a field experiment in a US public elementary school (Oak School). Teachers were deceived into believing that a set of one fifth of their class were expected to develop ("blossom and spurt") much faster than the rest, as measured by IQ points (supposedly measured by the

[30]The first self-fulfilling prophecy to be investigated extensively in psychology was the experimenter effect. The experimenter effect refers to the possibility of the experimenter influencing subjects to respond to the treatment in a way that conforms to the experimenter's expectations. Rosenthal (1963) summarized a dozen experimenter-effect studies and wondered whether similar interpersonal expectation effects occur among physicians, psychotherapists, employers, and teachers.

Harvard test of inflicted acquisition). In fact, this set was randomly selected. The main outcome measure was an IQ test (Test of General Ability), administered at the start of the school year (pretest) and at 4 months (end of first semester), 8 months (end of second semester and of first year of school), and 20 months (end of second school year with a different teacher). Rosenthal and Jacobson showed that the students for whom teachers had raised expectations had faster IQ gains than control students in the same classes (the treatment children gained 12 IQ points over the course of the year, and the control children gained 8), with the biggest effect on first and second grade children by the end of the first year. Rosenthal and Jacobson dubbed this boost in achievement driven by teachers' beliefs the "Pygmalion effect."

The Pygmalion effect in the classroom was subsequently studied intensively,[31] and criticized extensively. Snow (1995) re-analyzed the data from the original experiment, highlighting that approximately 35 percent of the IQ observations fall out of the normal range and that there are several observations which have rapid growth in pre-test and post-test scores (e.g. increasing from 17 to 110). He finds that the expectancy effect on IQ disappears when these outlier scores are omitted. A relatively recent review of the literature (including a balanced review of the meta-analysis and the various critics) concludes that while the Pygmalion effect in the classroom is real, it is probably fairly modest (Jussim and Harber, 2005).

Since this early work, social psychologists have demonstrated the self-fulfilling nature of leaders' expectations in several other field settings and have tried to better understand the underlying mechanisms. Rosenthal (1994) and Eden (1992) provide a review of much of the work in this literature. For example, Eden and Shani (1982) replicated the original design and results of Rosenthal and Jacobson (1968) in the Israeli Defense Forces. But they also concluded, based on additional survey work to complement the randomized controlled trial, that leadership behavior was a key mediator in generating the Pygmalion effect.

Also using the Israeli Defense Forces as a field, Eden and Ravid (1982) interestingly combined expectancy and self-expectancy manipulations in a single study. Trainees included 60 men in the first half-year of military duty enrolled in a seven-week clerical course divided into five training groups, each instructed by a commander. To produce the Pygmalion effect, a random quarter of each instructor's trainees were described to the instructor as having high success potential. An-

[31] See Dusek, Hall, and Meyer, 1985, Rosenthal and Rubin, 1978, which was one of the first meta-analysis in psychology and was based on 345 studies, and Rosenthal, 1994, which updates it for reviews

other random quarter were told directly by a psychologist in a brief personal interview that they had high success potential, in order to induce high self-expectancy. The remaining trainees served as controls. Learning performance was significantly higher in both high expectancy groups than in controls, confirming the Pygmalion hypothesis and the additional hypothesis that inducing high self-expectations similarly enhances trainee performance. Interestingly, while several instructors were unexpectedly relieved midway through the course, the hypothesized performance differentials continued even though the authors abstained from refreshing the expectancy induction among the substitute instructors, reflecting the possible durability of expectancy effects. Finally, Eden and Ravid (1982) also showed that equity considerations among the trainees likely played a mediating role: trainees in both of the high expectancy conditions reported feelings of over-reward, which may have motivated them to improve their performance.

While the Pygmalion literature shows that the self-fulfilling nature of raising leaders' expectations, another branch of this literature also demonstrated the self-fulfilling nature of lowering those expectations. Psychologists have dubbed this the "Golem effect".

There have been far fewer studies on the Golem effect than the Pygmalion effect, given the trickier ethical issues associated with lowering leaders' expectations (Reynolds, 2007). This challenge has led to research designs that are not quite as "clean" as those used to demonstrate the Pygmalion effect. For example, Oz and Eden (1994) randomly led treatment-assigned squad leaders (n = 17) in a military unit to believe that low scores on physical fitness tests were not indicative of subordinates' ineptitude, while control squad leaders (n =17) were not told how to interpret test scores. Tests indicated that low-scoring individuals in the experimental squads improved more than those in the control squads. While the researchers employed a respectable research design and were cautious to abide by ethical standards, the sample was extremely small and the researchers did not directly attempt to lower supervisors' expectations.

Given the challenge of doing research on the Golem effect, an alternative approach in the literature has been to rely on natural variation in leaders' expectations, rather than exogenously varying those expectations. For example, Babad, Inbar, and Rosenthal (1982) studied expectation effects among physical education student-teachers. They found that pupils about whom they imparted high expectations to the instructors performed best (i.e. the standard design for a demonstration of the "Pygmalion effect"). However, they also found that pupils toward whom instructors naturally harbored low expectations performed worse than those regarding whom

they had high or intermediate natural expectations, consistent with a "Golem effect."

A recent paper (Kondylis et al., 2015) demonstrates the power of self-fulfilling prophecy. In villages, either women or men were randomly selected to learn a new technology and teach it to others. Women retained more information from the training, and those who were trained by them did in fact learn more. But women ended up performing much worse in terms of the number of farmers they convinced, because other farmers perceived women as less able, and hence paid less attention to their messages.

2.2.2 Endogenous Responses to Bias

While the Pygmalion and Golem effects demonstrate the self-fulfilling nature of leaders' expectations about performance on that performance, they are not directly tied to discrimination. Are leaders' biases against some groups also endogenously affecting the performance of these groups? Two recent field studies in the economics literature provide what we believe is the first field-based answers to this question. Conceptually, these studies follow a very similar research approach to that in Babad, Inbar, and Rosenthal (1982) to demonstrate the relevance of self-fulfilling prophecies as an explanation for persistent differences in performance between different groups of workers or students. Specifically, rather than "artificially" priming leaders to vary their level of bias, the analysis relies on randomly assigning those trainees to leaders who are known to have different levels of bias. To be clear, the limit of this design compared to the preferred "Pygmalion design" is that any unobserved factors that are systematically correlated to different levels of biases among leaders cannot be formally ruled out as an explanation for the findings. However, the two papers below take several ingenious steps to deal with this concern.

Glover, Pallais, and Pariente (2015) studied cashiers in a French grocery store chain, a sizable share of whom were of North African and Sub-Saharan African origin. They assess whether cashiers performed worse on the days when they were assigned to a manager who was more biased against their group. They measured managers' bias towards workers of different origins using an IAT test. The cashiers in these stores worked with different managers on different days and had virtually no control over their schedule, allowing the authors to use the quasi-randomness of the schedules to assess the causal effect of being paired with a more biased manager. To address the difficulty raised above of manager bias being correlated with some other manager characteristics that might also affect employee performance (for example, more biased managers might also be

less skilled), they use a difference-in-difference methodology, comparing the change in minority workers' performance under biased and non-biased managers with the change in non-minority workers' performance under these two types of managers. They find that on days when they are scheduled to work with biased managers, minority cashiers are more likely to be absent. When they do come to work, they spend less time at work; in particular, they are much less likely to stay after their shift ends and they scan articles more slowly and take longer between customers.

Glover, Pallais, and Pariente (2015) also report interesting complementary survey evidence to better understand mechanisms. They do not find that minority workers report that they dislike working with biased managers more, or that biased managers dislike them, or that biased managers make them feel less confident in their abilities. However, they do find evidence that biased managers put less effort into managing minority workers. Minority workers report that biased managers were less likely to come over to their cashier stations and that biased managers demanded less effort from them. Consistent with this, they find that the effect of manager bias grows during the contract, perhaps as workers may learn that they are not being monitored by biased managers.

Lavy and Sand (2015) estimate the effect of primary school teachers' gender biases on boys' and girls' academic achievements during middle and high school, as well as on the choice of advanced-level courses in math and sciences during high school. In particular, they tracked three cohorts of students from primary school to high school in Tel-Aviv, Israel. They measured teachers' gender biased behavior by comparing their average marking of boys' and girls' in a "non-blind" classroom exam to the respective means in a "blind" national exam marked anonymously. For identification, the authors rely on the conditional random assignments of teachers and students to classes within a given grade and a primary school. They compare outcomes for students that attended the same primary school but were randomly assigned to different teachers, who have different degrees of stereotypical attitudes. They find that being assigned to a more gender-biased teacher at early stage of schooling has long-run implications for occupational choices and hence likely subsequent earnings in adulthood. Specifically, teachers' over-assessment of boys in a specific subject has a positive and significant effect on boys' achievements in the national test on that subject administered during middle and high school, while it has a significant negative effect on girls'. In addition, assignment to primary school math teachers favoring boys over girls encourages boys and discourages girls from engagement in advanced math courses offered in high

school.

2.3 Discrimination in Politics and Inequality across Groups

A direct consequence of discrimination in politics and other leadership positions is that there are fewer members of the discriminated group with the power to act in the interests of others in their group. In a standard median voter world, the underrepresentation of women or other subordinate groups in politics would matter less, as elected politicians would endeavor to represent the interest of the median voter. But if politicians cannot commit to a particular political platform, and their group membership eventually determines the type of policies they will implement, then the lack of representation at the top means that the underrepresented groups in society will get worse outcomes (Besley and Coate, 1997; Pande, 2003). This would also occur if the absence of a leader means that the underrepresented groups find that they cannot express their preferences in the political arena.

The best evidence on the consequences of discrimination in politics comes from studies that have evaluated what happens to the underrepresented groups when they finally gain political representation. A few observational studies have exploited exogenous shocks to representation due to close elections; a few other papers have also studied non-randomized variation in mandates.[32] There have also been a set of studies that exploit the random selection of places that have to elect a leader from a historically underrepresented group (caste, tribe or gender) in India's local governments. Comparing villages that were randomly selected to receive either a male or female head, Chattopadhyay and Duflo (2004) find that female leaders spend more on goods that women prefer, compared to those that men prefer. Beaman et al. (2010) replicate the results over a longer time period. Using a dataset that covers a larger number of states, they find that the results persist over time, and that investments in drinking water (a preferred good for women) continue to be higher even after the seat is not reserved anymore and women have (generally) left power.[33] Iyer et al. (2012) show that greater female representation (in local governments) is related to more crimes against women; using a household-level crime victimization survey in Rajasthan, they however show that the increase is not due to an actual increase in the amount of

[32] See Pande (2003), Clots-Figueras (2009), Clots-Figueras (2011), and Rehavi (2007) for examples.

[33] Bardhan, Mookherjee, and Parra Torrado (2010) compare places before and after reservation and do not find a difference in what leaders do, but since there seems to be a lingering effect of quota on pro-female policies, this finding might not be so surprising.

crimes, but rather greater willingness to report such crimes. Finally, Chattopadhyay and Duflo (2004) further find that village leaders from scheduled castes invest more in scheduled castes hamlets.[34]

2.4 Benefits of Diversity?

Discrimination leads to less-diverse firms, legislative assemblies, etc., but does diversity in itself matter for society? What are the implications of the low diversity that discrimination may generate for the performance of organizations and society in general?

2.4.1 Does Homogeneity Hurt or Help Productivity?

A long literature in political economy and development has tended to emphasize the *cost* of diversity, in particular ethnic diversity. If members of different groups do not like each other, diversity creates hold-ups, breeds conflicts, makes it difficult to agree on public good provision, etc.

Lang (1986) proposes that the roots of discrimination are communication difficulties across different groups (what he calls "language communities"). A similar argument is made by Lazear (1999). In that view of the world, segregation arises naturally, because homogenous groups are more productive (since communication within them is faster and easier). More homogeneous groups will create a trusting environment where people can work better together. While the minority will suffer as a result, the short-run equilibrium is efficient, and policies directly aimed at increasing diversity would be socially counterproductive. The role of policy should be instead to diminish language barriers between groups (through the education system, for example).

Others have emphasized the benefits of diversity, and potential drawbacks of "homophily" (or the tendency to want to associate only with people like oneself). One powerful argument is that similar people will tend to have similar information and perspectives, and if people only interact with people like themselves, lots of valuable information will be not transmitted across groups. Arguments along these lines have been made, more or less formally, in the human resources and management literature. More formally, Golub and Jackson (2012) show that, when agents in a

[34]Dunning and Nilekani (2013) find little impact of the reservation on distribution of goods by ethnic group and a strong impact of parties, but they use a regression discontinuity design strategy rather than focusing on a states where the assignment is random.

network prefer to associate with those having similar traits (homophily), it may take a very long time for participants in a network to converge to a consensus.

Ultimately, there is thus a trade-off between the cost of communication and collaboration and the benefits of diverse view points, which means that diversity (and hence homophily) may in theory hurt or improve productivity (Hamilton, Nickerson, and Owan, 2003; Alesina and Ferrara, 2005).

While there is a large non-experimental literature on the impact on diversity on public good provision[35] and a sizeable lab experimental literature,[36] the field experimental literature is more nascent. There are nevertheless a few interesting recent papers that we review below.

Hoogendoorn and Van Praag (2012) and Hoogendoorn, Oosterbeek, and Van Praag (2013) experimentally varied the composition of teams of undergraduate student required to start a business venture as part of a class. In teams of 12, students start up, sell stock, and run a real company with a profit objective and shareholders for a year. They ran the experiment on 45 teams and 550 students.

The composition of teams was varied by gender (men only, women only or mixed) and ethnicity (the fraction of non-Dutch ethnicity varied from 20 percent to 90 percent). Students then had a year to choose their venture, elect officers, conduct meetings, produce, sell, make money, and liquidate. This was a field experiment; the program played out over a year, and the incentives to do well were very high. Students' ability to graduate, their grades, and potentially some money, were all on the line.

There is a clear benefit to gender diversity in the experiment. The performance as a function of the share of women in the team is inverse U-shaped, with the peak reached when the share of women is approximately 0.55. The authors attribute this effect to greater monitoring in gender-diverse groups. This in itself is an interesting finding as this is not a mechanism that is emphasized in the theoretical literature: perhaps when communication is too easy, workers become more complacent.

For ethnic diversity, the result is more subtle: Hoogendoorn, Oosterbeek, and Van Praag find that the marginal effect of increasing diversity on performance is zero or perhaps even negative when the teams are least 50 percent Dutch. However, once the teams are less than 50 percent

[35] See Alesina and Ferrara (2005) for a review of the literature on diversity

[36] For example, see Woolley et al. (2010) and Engel et al. (2014)

Dutch, further increases in the share of other groups are associated with better performance. The authors also identify evidence for the different channels proposed in the theoretical literature (including higher communication costs in more diverse groups, but also more diverse knowledge in more diverse teams), but these results are not extremely precise.

Hjort (2013) analyzes a natural experiment where a flower firm in Kenya randomly assigned workers to teams. Kenya offers a context with heightened ethnic tensions, and where the level of distrust among different groups may be particularly high. In the experiment, an upstream worker distributed flowers to a team of two downstream workers. The upstream worker earned w per flower packed, and the downstream worker, $2w$ per flower packed. Hjort finds that, conditional on productivity, upstream workers distribute fewer flowers to teams when one or both are not from his ethnic group, at the cost of lower wages for him, and lower production overall. Furthermore, within mixed teams, upstream workers give more flowers to the worker from their same ethnic group. Interestingly, the output gap between homogenous and ethnically mixed teams doubled during the period in 2008 when ethnic conflict intensified. In response to this, the firm introduced team pay for the downstream workers (not randomized) and subsequently experienced an increase in the productivity of the ethnically mixed teams.

Also in Kenya, Marx, Pons, and Suri (2015) randomly assigned enumerators to pairs, and each pair to a supervisor. The job of the enumerator was to make contact with a household and administer an intervention. They find that homogenous pairs have higher productivity, and they attribute that to higher trust in those teams. However, when a pair is further matched with a supervisor of the same ethnic group, the productivity is lower (not higher). The contrast between the (negative) impact of diversity in horizontal teams, and the (positive) impact in vertical relationships in the Marx, Pons, and Suri (2015) experiment hints at a different potential negative impact of discrimination: in-group preference may create room for cheating and corruption. In their setting, the co-ethnic supervisor was willing to let the enumerators cheat.

2.4.2 Discrimination and Corruption

Prendergast and Topel (1996) provide a theoretical analysis of the influence of favoritism on optimal compensation and extent of authority for the manager. The point extends further than the firm: discrimination may lead to misallocation of resources by politicians (to members of their ethnic group), or conversely to willingness to put up with corrupt or incompetent politicians

from one's own group (rather than less corrupt ones from another group) (Key, 1949; Padro i Miguel, 2007). More generally, voters' preferences for a group may diminish the role of issues in campaign, and by implication the quality of government (Dickson and Scheve, 2006).

Besley et al. (2013) provide non-experimental evidence of this effect: they show that in Sweden, after the social democratic party imposed gender balance by requiring that all candidates be selected in a "zipper" pattern (one man/one woman), the quality of the male candidates greatly increased (they call this the "crisis of the mediocre man").

Experimental evidence of the link between homophily and the quality of politicians or corruption level is rare. One interesting experiment took place in Uttar Pradesh, India's most populous state, where the rise in caste-based politics has been accompanied by a staggering criminalization of politics (Banerjee et al., 2010). On the eve of the 2007 election, 206 of the sitting members of the legislative assemblies had a criminal case pending against them (Banerjee and Pande, 2009). Prior to the 2007 election, the authors conducted a field experiment in which villages were randomly selected to receive non-partisan voter mobilization campaigns (street plays, puppet shows, or discussions). One type of campaign encouraged citizens to vote on issues, not on caste, while the other encouraged them to not vote for a corrupt candidate. They found that the caste campaign led to a reduction of the (reported) votes on caste, and to a reduction in the vote share going to candidates with a criminal record. It thus seems that successfully reducing discrimination (in this case, to be more precise, reducing lower caste group members' tendency to systematically discriminate against higher caste candidates) does lead to an improvement in the quality of elected leaders.

The natural experiment in India discussed in Section 2.3 that introduced quotas for women in politics shed interesting light on this question as well. In the short run at least, reservation for women politicians reduced bribe taking (Beaman et al., 2010). Of course, in the short run, quotas do not increase competition (since on the contrary the pool is reduced to women only, whereas it was initially open to women and men) and the observed reduction of corruption could be due to inherent characteristics of women, or to their lack of experience. However, quotas do tend to *increase* political competition in the medium run, because once a woman leaves office and her seat is open, she (or her relative) have the option to run again, but the field is now more open to competition than if she were a traditional incumbent. Moreover, when Banerjee et al. (2013) collected data on what happens in previously reserved places, they found that female

incumbents whose seats became free were less likely to run than male incumbents whose seat became free, but that this effect disappeared when they considered not only the incumbent, but the incumbent and his or her family. In other words, the probability to elect someone from the incumbent's dynasty remains the same in places that just have experienced a quota or not. Also, they found that the probability of re-election of someone from the incumbent's family is more sensitive to past performance in the previously reserved places. Thus, the best politicians' dynasties are more likely to be re-elected after reservation, and the worst ones less likely to. To the extent this effect persists, it does suggest that policies that constrain voters to vote outside of their "comfort" zone may improve the quality of the decision-making process overall even after these constraints are lifted.

2.4.3 Law of Small Numbers

Even if discrimination does not lead to outright corruption, it may restrict the pool of available candidates. Research shows that the leader quality matters both for firms (Bertrand and Schoar, 2003) and for countries (Jones and Olken, 2005). If discrimination implies that leaders have to be selected from a relatively small pool, it reduces the chance that the most talented person will be picked, and it thus may have negative productivity consequences.

The empirical evidence (even non-randomized) on any such consequence of discrimination is thin at best: Ahern and Dittmar (2012) and Matsa and Miller (2013) examine the impact of the Norway 2006 law which mandated a gender quota in corporate board seats. They both find negative consequences on profitability and stock prices. However, these are short-run impacts. It could be that women are temporally less effective because they are less experienced, or that they maximize something else other than short-run shareholder value, which may turn out to be profitable in the long run.

Unfortunately, we don't see an experiment on this, nor can we think of an obvious design for one. But this would be a very intriguing avenue for further research.

3 What Affects Discrimination?

3.1 Leaders and Role Models

One effect of discrimination against a group is that few leaders emerge from it into the mainstream. This has potentially three consequences. First, mechanically, fewer people from this group are in a position to make a decision regarding others. To the extent that leaders discriminate against members from other groups, discrimination will persist. Second, the majority group may be reinforced in their belief that the minority group is incapable of success, since they have rarely, if ever, observed success of the minority in practice. And third, members of the minority group may then feel that either they are incapable of succeeding, or that the world is rigged against them, and there is no point in even trying.[37] Given all this, discrimination could be lessened by forcing exposure to leaders from groups that are traditionally discriminated against, which is often achieved through quotas. This section reviews the evidence for this and also points out the gaps.

3.1.1 Does Diversity in Leadership Positions Directly Affect Discrimination?

Mechanically, discrimination may breed discrimination, because the decision-making power is concentrated with the majority group. For example, if managers are mostly males, they may tend to favor other males in their recruiting or promotion decisions. This may happen because they themselves discriminate (consciously or unconsciously) or because they know more males and are more likely to promote or hire people they know or who are more similar to them.[38] This tendency is part of the rationale for requiring a certain fraction of women on corporate boards, in academia, or in appointment, evaluation and promotion committees.

It is, however, not obvious that minority group leaders, or committees that contain such minority leaders, would necessarily favor others from the minority: faced with their own discrimination, they may feel the need to go to lengths to avoid being perceived as biased. In several observational studies, women were not inclined to judge other women more favorably than men.[39]

[37] See for example Lockwood and Kunda (1997)

[38] Bagues and Perez-Villadoniga (2012) provide evidence from entry exams into the judiciary in Spain that support the latter effect; Bagues and Zinovyeva (2015) show that the former effect also applies in the case of academic promotions.

[39] See Booth and Leigh (2010) for an audit study in Australia, where they find no interaction between the gender on the résumé and the gender of the recruiter. See also Broder (1993) for similar evidence in the context of NSF proposal reviews, and Abrevaya and Hamermesh (2012) for referee reports.

In group decisions, there may also be a response of other members of the committee, who may try to "undo" any agenda they perceive (rightly or wrongly) the minority group members to have.

The empirical evidence of the impact of minority representation on selection committees largely comes from a series of very interesting papers by Bagues, Zinovyeva, and their co-authors. Bagues and Esteve-Volart (2010) examine the impact of the gender composition of the evaluation committee for the entry exam in the Spanish judiciary on the success of women in that exam. A causal study is made possible by the fact that people are randomly assigned to a committee. They find that women are *less* likely to succeed at the exam when the committee they are assigned to has more women, while the opposite is true for male candidates. Additional evidence in the study suggest that these results might be driven at least in part by the fact that female evaluators tend to overestimate the quality of male candidates.

Zinovyeva and Bagues (2011) and Bagues, Sylos-Labini, and Zinovyeva (2014) present interesting evidence from randomized academic evaluation committees in Spain and Italy, respectively. In both countries, candidates for promotion appear in front of a centralized committee to be qualified. Files are assigned to randomly composed committees. In the Spanish case, the authors find no impact of an additional female committee member on the promotion likelihood of female candidates. In the Italian case, they find a *negative* effect: in a five-member committee, with each additional female member added to the committee, the success rate of female applicants relative to that of male applicants decreases by around two percentage points. Analyzing the voting records, they find both that (1) the same female candidate is scored on more harshly by females than by males, and (2) male committee members grade female candidates more harshly when there are women on the committee, perhaps because they are trying to compensate for a perceived bias in favor of women on such committees (even though in reality the opposite appears to be true given (1)).

This evidence on academic and recruitment committees is striking and suggest that some type of affirmative action may in fact hurt promising female candidates. It would be interesting to see if it also carries through in other settings, such as management or political decisions. Bursell (2007) is an audit study that makes some progress in this direction (although the specific comparison it focuses on is itself not experimental). He sent 3552 applications to 1776 jobs in Sweden, including applications to more skilled positions, such as senior/high school teachers, IT-professionals, economists, and engineers, and compares, among other things, the call-back rates

for applicants with Swedish-sounding and non-Swedish-sounding names according to the name of the CEO. He finds, consistent with the evidence above, that when the "CEO of a company has a foreign sounding name, the applicants with a Swedish sounding name have a 2.4 times higher probability to receive a call-back. If the CEO has a Swedish sounding name, the probability is 1.7 times higher" (Bursell, 2007).

3.1.2 Minority Leaders and the Attitude of the Majority

Even if there is no direct effect of having women or minority members in on leadership decisions, it could still affect discrimination against the minority because those minority individuals in leadership positions will change the beliefs, or precision of the beliefs, of the majority about the competence of the minority group.

In a working paper version of Beaman et al. (2009), the authors propose a model where taste and statistical discrimination reinforce each other. Suppose that there are strong tastes (or social norms) against having a female leader. Then, it is very likely that citizens have never observed a female leader in action. This makes female leaders riskier as a group: even if citizens believe that female leaders are equally competent on average, they have much more precise priors about male leaders, and to the extent they are risk averse, this will lead them to avoid women leaders. This is of course reinforced if citizens start with the prior that women are less competent: they will never have the occasion to find out that in fact they are wrong. In this world, forcing exposure to minority leaders (political leaders, board members, colleagues in academic departments, students at top colleges, etc.) will have a persistent negative impact on discrimination, even if it does not affect the underlying taste for the community, simply by affecting beliefs about the competence of the minority group.[40]

The impact might be reinforced if the image of what constitutes a good leader also evolves in response to what people see. Eagly and Karau's (2002) "role incongruity" theory stipulates that one reason why people prefer male leaders is that the traits associated with leadership (strength, assertiveness) are not traits that are associated with women under prescriptive gender norms (such as being nice, accommodating, etc.). Yet, as people get to see many (not just one or two

[40] In an observational study, Miller (2014) finds evidence consistent with such effects in the context of affirmative action programs in the US. US government contractors are forced to hire minority workers. Miller finds that, after an establishment is no longer subject to such affirmative action because it stops being a government contractor, the black employment share nevertheless continues to grow.

token) women leaders who are strong, but also nice, or who have an effective, but also more accommodating, leadership style, then people may change their attitude towards female leaders. As inconsistencies between the female gender stereotype and qualities associated with leadership diminish, so will prejudice towards female leaders.

Of course, a potential force in the opposite direction is the possibility of backlash against minority leaders, in particular if there is a perception that they got into their leadership role because of special treatment (Coate and Loury, 1993).

Beaman et al. (2010) study a natural experiment in the context of local electoral politics in India, and are able to provide more evidence for the mechanism underlying the persistent effect of temporary affirmative action policies. In a context where local village councils are randomly selected, by rotation, to be forced to elect female leaders, the authors show that after a cycle of reservation (and even more when the same place happened to be reserved for a woman for two cycles in a row), more women run, and are elected, on un-reserved seats. While there could be many reasons for this (including the fact that women may have become more willing to run, or that networks of women may have been created), they provide evidence that it is probably at least in part due to a change in attitude in the villages that were previously subjected to reservation. They collect evidence on attitudes in various ways: with a Goldberg-type experiment, and with two IATs, one for like or dislike for female leader (a more "hardwired" attitude that their model takes no stance on) and one for a stereotype associating women with domestic activities and men with leadership activities (in the spirit of the assessing whether "role incongruity" effects diminished). They find that the experience with the past quota does not affect preferences (as measured by the taste IAT), although it tends to harden stated preferences against women in leadership. However, citizens (particularly men) update on measures of perception of women's competence. For example, their rating of a speech pronounced in female voice converges to that of a speech given by a male voice if they have been exposed to a quota either in this cycle or in the previous cycle. Moreover, the stereotypical IAT also shows a decrease of the stereotype that associates women with domestic activity rather than with leadership. This provides reasonably strong field evidence that exposure to role models from another group affect attitudes. The evidence seems quite robust: Bhavnani (2009) also finds that women continue to be more likely to be elected after a seat was reserved for a while (in urban Maharashtra). Banerjee et al. (2013) also find, in Rajasthan, that women are more likely to run (and win) on a previously reserved

seat.

Although there is a vast laboratory literature that test the role-incongruity theory and its implications, and laboratory studies that show, for example, that college students asked to screen candidates for a typical male job (e.g. finance manager trainee) are less likely to discriminate against a female résumé after reading an editorial documenting women's success in this type of job (Heilman and Martell, 1986), we are not aware of field experiments that investigate these types of effect in other contexts (e.g. exposure to a female or black manager, minority teachers, etc.). It would be valuable to establish whether such impact on the majority's attitude can also be documented in some of these other contexts.

3.1.3 Role Models, Aspirations, and the Attitude of the Minority

Leaders issued from disadvantaged groups, in addition to their direct decision-making power and to the effect they may have on the opinion of the majority could also serve as role models and trailblazers. They might affect the attitudes of the minority group about their own ability to succeed, or their aspirations to do so. Seeing successful women or Blacks may lessen stereotype threat (as discussed in Section 3.1.2), or the belief among those groups that society is rigged against them so there is no point in trying anyway. In both cases, exposure to role models may increase effort and lead to better outcomes for the minority, even without direct changes in the majority attitude (though this could of course trigger subsequent change in the majority's beliefs and attitudes as well). However, as noted by Lockwood and Kunda (1997), these positive effects might be moderated by how fixed minority group members view their ability, and how personally relevant and attainable they consider the achievement of the role models.

As in the case of the impact of exposure on the attitudes of the majority, there is both a descriptive and a laboratory literature on this question. The observational literature looks for correlation between either outcomes (e.g. teen pregnancy) or measure of stereotyping (e.g. IAT) to naturally occurring exposure (e.g. black teachers). For example, Dasgupta and Asgari (2004) show that women who have been exposed to female teachers and role models are more likely to associate women and leadership. A laboratory experiment literature explores the extent to which exposure to stereotypically feminine role model in a science, technology, engineering and math (STEM) career (Betz and Sekaquaptewa, 2012) or exposure to a non-stereotypical computer science role model (Cheryan et al., 2011) increases the likelihood that girls will present

themselves as interested in STEM. Interestingly, and maybe in line with the cautionary note in Lockwood and Kunda (1997), these studies show that a role model that simply belongs to the minority group might not be sufficient, and that the "type" of role model appears to matter significantly.

Cheryan et al. (2011) find that exposure to a non-stereotypical computer science role model (e.g. someone who dresses fashionably, enjoys sports and hanging out with friends, and watches "normal" TV shows, such as *The Office*) through a "getting to know each other" task in the lab increases female subjects' beliefs of succeeding in the field, and that this is true regardless of the gender of that role model. Similarly to the discussion in Section 2.1.2 on "social identity," the authors argue that this is because women feel the stereotypical characteristics of a computer science major (e.g. social isolation, obsession with computers, social awkwardness) do not align with what they see as their female gender role (e.g. helping others, having social skills, attending to physical appearance), and a non-stereotypical role model, whether man or woman, can thus influence young girls' preferences and beliefs. On the other hand, Betz and Sekaquaptewa (2012) find that counterstereotypic-yet-feminine STEM role models (as signaled by characteristics such as wears make-up and pink clothes, enjoys reading fashion magazines, etc.) discourage middle school girls' success expectations in STEM relative to gender-neutral STEM role models (as signaled by characteristics such as wears dark-colored clothes and glasses, enjoys reading books, etc.). The authors find that this is particularly the case for girls who did not identify with STEM subjects, and conclude that this subgroup of girls viewed the combination of both STEM-success and femininity as unattainable.

Reviewing either literature fully is outside the scope of this chapter, but the mixed results of the lab experiments provide interesting directions for field research..

Here again, however, the field experiment work so far appears to be quite limited. Beaman et al. (2012) study the same randomized natural experiment for women in leadership positions in India, and look at the impact on girls' educational attainment and career aspirations. They give evidence of impact on parents' hopes for daughters. Compared to never-reserved villages, parents in reserved villages were more likely to state that they would like their girl to graduate or study beyond the secondary school level, and more likely to state that they would like their daughter to have a career. Parental aspirations for boys did not change. Furthermore, in villages with reservation, girls were more likely to stay in middle school, which cannot be directly attributed

to any direct action of the leader because middle schools are not under their jurisdiction. This is therefore strongly suggestive of a causality running from role model to change in aspirations and actual change in behavior.

Overall, this literature seems to us surprisingly thin, compared to the larger literature on "horizontal exposure" (e.g. roommates or classmates) which we discuss below in Section 3.2. Part of the explanation for this is practical: there is probably more naturally occurring variation in peer groups than in supervisors, leaders or teachers. Another issue is that in many settings, female teachers or leaders may take actions that can translate directly into behavioral changes for female students (or trainees) even absent any effect on aspirations, so that isolating the impact on minority aspirations is tricky. While this was not the case in the quota experiment in India, it could have been. Nevertheless, we suspect that the lack of more field studies in this area is also a reflection of too little attention devoted to this important and exciting topic, and that much more probably can be done to explore how exposure to role models affects minority groups' aspirations.

3.2 Intergroup Contact

Allport (1954) is often credited with the development of the contact hypothesis, also known as Intergroup Contact Theory. The premise of Allport's theory is that, under appropriate conditions, interpersonal contact is one of the most effective ways to reduce prejudice. The theory, which was originally devised for encounters across racial and ethnic lines, states that if majority group members have the opportunity to communicate with minority group members, they are able to understand and appreciate them. As a result of this new appreciation and understanding, prejudice should diminish. Allport's proposal was that properly managed contact between the groups should reduce prejudice and lead to better interactions. In particular, Allport held that reduced prejudice will result when four features of the contact situation are present: equal status between the groups in the situation, common goals, intergroup cooperation, and the support of authorities, law, or custom.

Much of the psychology literature on the contact hypothesis has focused on lab experiments that have helped refine Allport's original theory. An unresolved issue in psychology is whether specific conditions for the contact situation are needed to ensure that contact will have the theorized effect. For example, is it important for the contact to take place in a cooperative

environment with peers of equal status (e.g. two roommates in a university dorm working together on a math homework) for contact to be effective at reducing discrimination? In a meta-analysis combining observational and experimental studies of the intergroup contact theory, Pettigrew and Tropp (2000) find that intergroup contact reduces prejudice in 94 percent of the 515 studies reviewed. Their meta-analysis also suggests that the contact effect generalizes to a broad range of minority groups (not just racial and ethnic minorities but also the elderly, the mentally ill, LGBT, etc.) as well as a broad range of contact settings (schools, homes, etc.). Pettigrew and Tropp (2000) also assess whether the optimal conditions for contact stated by Allport are necessary for positive contact outcomes. They find that the inverse relationship between contact and prejudice persists, though not as strongly, even when the contact situation is not structured to match Allport's conditions. Hence, while Allport's conditions may not be necessary for prejudice reduction, some combinations of them might be relevant. Psychologists are still debating and investigating the specific negative factors that may prevent intergroup contact from diminishing prejudice.

While much of the experimental research on the contact hypothesis has taken place in the lab, there have also been quite a few field experiments. Green et al. (forthcoming) identify 56 field experiments.

The best known field work within economics has focused on contact between college roommates. Sacerdote (2000) was the first to exploit the random assignment of roommates at college for a study of peer effects on test scores. More relevant to us, Boisjoly et al. (2006) leveraged random roommate assignment at Harvard to study the impact of shared experiences at college on opinions about the appropriateness of keeping affirmative action policies. They find that white students who are randomly assigned African-American roommates are significantly more likely to endorse affirmative action. Hence, mixing with African-Americans tends to make individuals more empathetic to them. They also find that white students who were assigned roommates from any minority group are more likely to continue to interact socially with members of other ethnic groups after their first year.

What remains unclear from Boisjoly et al. (2006) is whether contact to a minority roommate reduced stereotype or bias. Empathy might increase even if bias is unaffected. Burns, Corno, and La Ferrara (2015) leveraged the same design to get at this question, and hence their paper is closest to a field test of the contact hypothesis. Specifically, they exploit random assignments

of roommates in double rooms at the University of Cape Town to investigate whether having a roommate of a different race affects inter-ethnic attitudes, but also cooperative behavior and academic performance. They find that living with a roommate from a different race significantly reduces prejudice towards members of that group, as measured by an IAT. The reduction in stereotype is accompanied by a more general tendency to cooperate, as measured in a Prisoner's dilemma game, but smaller effects on trust, as measured in a trust game. The paper also reports interesting results on grades. Black students that are assigned a non-black roommate experience higher GPAs; white students that are assigned a non-white roommate experience lower GPAs.

Related findings are reported in Shook and Fazio (2008). Participants were white freshmen who had been randomly assigned to either a white or an African-American roommate in a university college dormitory system. Students participated in two sessions during the first two and the last two weeks of their first quarter on campus. During these sessions, they answered questions about their satisfaction and involvement with their roommates and completed an inventory of intergroup anxiety, as well as an IAT test. Automatically activated racial attitudes (as measured with the IAT) and intergroup anxiety improved over time among students in interracial rooms, but not among students in same-race rooms. However, participants in interracial rooms reported less satisfaction and less involvement with their roommates than did participants in same-race rooms.

Several field experiments in psychology have also examined contact effects between classmates. In particular, psychologists have looked at cooperative learning techniques – which are designed so that students must teach to one another and learn from one another and place a strong emphasis on the academic learning success of each member of the group – and tested whether these help reduce prejudice (Johnson and Johnson, 1989; Slavin, 1995). The rational is that because cooperative learning encourages positive social interactions among students of diverse racial and ethnic backgrounds, it creates some of the conditions hypothesized in Allport as beneficial to reducing discrimination: as students work cooperatively, they have the opportunity to judge each other on merits rather than stereotypes. Slavin and Cooper (1999) provide a review of the field evidence on cooperative learning, which has been generally supportive of cooperative learning being a useful tool to improve inter-group relations.

For example, Slavin (1977) and Slavin (1979) study one particular approach to cooperative learning, called Student Teams Achievement Divisions. Under this approach, the teacher presents

a lesson, and students then study worksheets in four-member teams. Following this, students take individual quizzes, and team scores are computed based on the degree to which each student has improved over his or her own past record. The team scores are published in newsletters. He finds that the students who had experienced such cooperative learning over periods of 10 to 12 weeks gained more cross-racial friendships than did control students. In a follow-up one year later, Slavin (1979) found that the students that experienced cooperative learning named an average of 2.4 friends outside their own race, compared to an average of less than one in the control group. Also, Slavin and Oickle (1981) found significant gains in Whites' friendships toward African-Americans as a consequence of using the same cooperative learning method, but, interestingly, no difference in African-American friendships toward Whites.

A recent paper in economics also brings the contact hypothesis to the classroom, but under conditions that are not tailored to be as optimal as possible for prejudice reduction to occur, at least as hypothesized by Allport. Starting in 2007, some elite private schools in Delhi were required to offer places to poor students. Rao (2013) exploits this policy change and uses a combination of experimental and administrative data to study whether exposure of rich students (from 14 private schools in New Delhi) to poorer students affects (1) tastes for socially interacting with or discriminating against the poor, (2) generosity and prosocial behavior, and (3) learning and classroom behavior. Core to his identification strategy is a comparison of outcomes for treated and non-treated student cohorts within a school. Rao also exploits a second identification strategy that is closer to a randomized design. Some schools in his sample used the alphabetic order of first name to assign students to study groups and study partners. Hence, in those schools, the number of poor children with names similar to a given rich student provides plausibly exogenous variation in personal interactions with a poor student. This second identification strategy is obviously more appealing as a test of the contact hypothesis, because it focuses more centrally on changes in personal interactions between students, and rules out other confounds (such as changes in teacher behavior, changes in the curriculum, etc.).

Rao (2013) finds that economically diverse classrooms cause wealthy students to discriminate less against other poor children outside school. As discussed in Section 1.9, Rao's approach to measure discrimination is quite unique. First, he relies on a field experiment in which rich participants select teammates for a relay race and are forced to reveal how they trade-off more-athletic poor students versus less-athletic rich students. Using this measure of discrimination,

Rao finds that exposure to poor students at school reduces discrimination by 12 percentage points. Rao also conducts a second field experiment. He invites students to attend a play date at a school for poor students, and elicit incentivized measures of their willingness to accept. He finds that having poor classmates makes students more willing to attend the play dates with poor children. In particular, it reduced the average size of the incentive that is required to attend the play date by 19 percent. Having a poor study partner (e.g. contact alone) explains 70 percent of the increase in this "willingness to play."

When Rao (2013) turns to how exposure to poor students affects pro-social behavior and learning in the classroom, he finds that having poor classmates makes students more prosocial, as measured by their history of volunteering for charitable causes at school, as well as their behavior in dictator games conducted in the lab. The findings reveal that exposure to poor students does not just make rich students more charitable towards the poor; instead, it affects generosity and notions of fairness more generally. Finally, Rao shows that exposure to poor classmates has limited effects on the wealthy students' test scores: while he detects marginally significant but meaningful decreases in rich students' English test scores, he finds no effects on Hindi or Math scores, or on a combined index over all subjects.

The studies reviewed above suggest that that inter-group contact is an effective tool to reduce prejudice, even though more work remains to be done to ascertain the specific conditions under which contact will be most effective. Yet, some recent work in psychology (Dixon et al., 2012) suggests new angles through which the contact hypothesis should be evaluated, and more specifically, the possibility that its impact on the ultimate goal of achieving a more inclusive society might be less obvious than what would immediately appear. One of the observations made under this new line of work is that prior research on the contact hypothesis has paid little attention on how the minority group reacts to contact, with the focus being on how contact changes prejudice level among majority group members. In this context, Dixon et al. (2012) mention a few observational studies suggesting that, while majority group members may demand more social change towards inclusiveness subsequent to inter-group contact, minority group members may actually become less demanding of social change, as they perceive that discrimination and social injustice have lessened. A few recent studies (Saguy et al., 2009; Dovidio et al., 2009; Glasford and Calcagno, 2012) provide lab results consistent with this observational data, with minority group members under the contact condition appearing lulled into believing that the majority

group is more just-minded than it really is. If these effects are real, one can easily imagine how contact may backfire at the societal level, with the theoretically more powerful advocates for the minority group (for example, African-Americans at Ivy League Universities experiencing positive inter-group contact) decreasing their level of political activism. At the very least, this provocative new research in psychology suggests that future field work on the inter-group contact hypothesis should be more systematic in collecting evidence on how minority group members react to contact, and broadening the definition of a successful intervention outcome.

3.3 Socio-Cognitive De-Biasing Strategies

In the absence of direct contact, is it possible to *teach* individuals to become less biased against the minority group?

We start with the discussion of a field experiment in Rajasthan, India, which offers a cautionary tale about how easy it might be to simply tell people to overcome their stereotypes. Banerjee et al. (2013) set up a large-scale randomized experiment designed to test whether citizens can learn from others' experiences about the quality of female leaders. This is an environment where, we have already shown, there is a large bias against the ability of women to be decision-makers. Using high-quality street theater troupes, they set up a street play followed by a discussion of the performance of local leaders a few weeks ahead of the 2010 panchayat (local government) election. Following up on the work we discussed previously in Section 3.1.2, which showed that direct experience with a female leader does change attitude towards female leaders (and willingness to vote for one), this study sought to test whether the process could be accelerated by providing citizens objective information that, in fact, women and men are about equally good at carrying out a key task in the local government. The experiment took place in 382 panchayats: in randomly selected ones, a street play emphasized the importance of the local leader in making key decisions, and encouraged citizens to vote for a competent leader. It then showed information on the average performance of all leaders in providing employment under the flagship employment guarantee scheme. In another group of villages, the play and the information was almost the same, but the script of the play emphasized the fact that citizens are often biased against women leaders, but that women also can be good leaders. The statistics provided on leader performance were also disaggregated by gender (as it turns out, women do about as well as men in the sample districts).

There are two main results: First, the play and information campaign, when it does not emphasize gender, does appear to move priors. More candidates enter and the incumbent is less likely to enter and to win. For example, the incumbent vote share declines by 6 percentage points (or a remarkable 60 percent) in villages where the general campaign was run. Moreover, the vote share for the incumbent become more sensitive to past performance in places where the gender-neutral campaign was run. Second, however, these effects disappear in places where the campaign introduces the "gender" theme: in those villages, there is very little effect of the intervention on any outcomes (including on the probability that a female runs or wins, or on the vote share for women). It is as if, when citizens understood that the campaign was about convincing them to consider women, they lost interest. These findings underscore the challenge of fighting discrimination in an environment where discrimination is rife.[41]

It is possible that this experiment failed because it did not pay enough attention to the structure of the bias and ways to overcome it. Over the last 20 years, social psychologists have designed and tested in the laboratory setting a series of strategies to reduce bias and stereotypical thinking. These include (following the categorization in Paluck and Green (2009)): consciousness-raising, targeting emotions through perspective-taking, targeting value consistency and self-worth, expert opinion and accountability interventions, as well as re-, de- and cross-categorization techniques.

Consciousness-raising strategies are inspired by the large body of work (in particular the IAT literature) suggesting that prejudice can operate without the person's awareness or endorsement of it. The most promising consciousness-raising strategies emerging from the psychology literature to date include counter-stereotype training and approach-avoidance training.

For example, in Kawakami et al. (2000), lab subjects received extensive training in negating specific stereotypical thinking towards elderly people and skinheads (young individuals with closed-cropped or shaven heads who typically wear heavy boots, are often part of the working-class, and stereotypically perceived as aggressive). In the elderly stereotype negation condition, subjects were instructed to respond "NO" on the trials in which they saw a picture of an elderly person paired with an elderly stereotypic trait and "YES" when they saw a picture of an elderly

[41]Note that the effect of reservation in this sample on the probability that a woman runs or wins after the reservation is cancelled is still positive, as in West Bengal or Mumbai: so the results are not due to the fact that people in Rajasthan are so hell bent against women that they cannot learn about them. It just appears they cannot learn about them from this intervention.

person with a nonstereotypic trait. In the skinhead stereotype negation condition, subjects were to respond "NO" on trials in which they saw a picture of a skinhead paired with a skinhead stereotypic trait and "YES" on trials in which they saw a picture of a skinhead paired with a nonstereotypic trait. Kawakami et al. (2000) show that such training in negating stereotypes was able to reduce the stereotypical activation. These results were obtained even when participants were no longer instructed to "not stereotype," and, importantly, for stereotypic traits that were not directly involved in the negation training phase. This reduced activation level was still clearly visible 24 hours following the training session.[42]

Dasgupta and Greenwald (2001) report on two experiments where they examined whether exposure to pictures of admired and disliked exemplars can reduce automatic preference for white over black Americans and younger over older people. In Experiment 1, participants were exposed to either admired black (e.g., Denzel Washington) and disliked white individuals (e.g., Jeffrey Dahmer), disliked black (e.g., Mike Tyson) and admired white (e.g., Tom Hanks) individuals, or nonracial exemplars. Immediately after exemplar exposure and 24 hours later, they completed an IAT that assessed automatic racial attitudes and two explicit attitude measures. Exposure to admired black and disliked white exemplars significantly weakened automatic pro-white attitudes for 24 hours beyond the treatment but did not affect explicit racial attitudes. Experiment 2 provided a replication using automatic age-related attitudes. Also, Wittenbrink, Judd, and Park (1997) examined the effects of watching videos of African-Americans situated either at a convivial outdoor barbecue or at a gang-related incident. Situating African-Americans in a positive setting produced lower implicit bias scores.

Kawakami, Dovidio, and Van Kamp (2007a) perform another lab experiment on negating stereotypical associations but focus on outcomes that are closer to those we might wish to affect in the real world. Participants first underwent gender counter-stereotype training, by pairing male faces with words like "sensitive" and female faces with words like "strong." They next evaluated four applications (résumés and cover letters) for a position as "chairperson of a District Doctor's Association." All of the applicants were qualified, but two had male names and two had female names (counterbalanced so that half the participants saw a particular résumé with a male name and the other half saw that same résumé with a female name). The evaluation

[42]Two follow-up studies outside of Kawakami's lab have partially replicated but partially qualified the original findings. See Gawronski et al. (2008).

of applicants involved two separate stages: judging the applicants along 16 different dimensions (eight stereotypically masculine traits like "risk-taker" and eight feminine traits like "helpful") and then simply choosing the best candidate. Some participants made the trait judgments first and chose the best candidate second, while other participants completed the two tasks in the opposite order. Among participants who had received no training, only 35 percent chose a woman for the job. In contrast, among participants who had undergone counter-stereotype training, 61 percent chose a woman.[43]

Kawakami et al. (2007b) also found that participants can change their implicit biases and unreflective social behaviors through "approach-avoidance" conditioning. In this study, white and Asian participants repeatedly pulled a joystick toward themselves when they saw black faces and pushed it away when they saw white faces. In pulling the joystick in, it was as if participants were bringing the perceived image closer, or "approaching" it. This training significantly reduced participants' implicit bias on the IAT. The same "approach-avoidance" conditioning training has also been shown to be promising (in the lab) to deal with the stereotype threat. Kawakami et al. (2008) report the beneficial effects for female undergraduates of repeatedly "approaching" math-related images ("e.g., calculators, equations"). After the training, those who initially reported that they did not like math and were not good at it tended to identify with and prefer math on implicit measures, as well as to answer more questions on a math test. In a series of follow-up studies, Forbes and Schmader (2010) replicated Kawakami et al. (2008), but built a longer delay (24–30 hours) between the de-biasing training and the math test, and also compared the relative effectiveness of approach-avoidance training with counter-stereotype training. They found that gender-math counter-stereotype training seemed more effective than approach-avoidance training. Women trained to associate the phrase "women are good at" with math-related words exhibited increased working memory as well as improved performance on math questions from the GRE (a graduate-level standardized test).

Because emotional states can influence the expression of prejudice, psychologists have hy-

[43]Interestingly, these effects were only observed when the task of choosing the best candidate came second, after the trait evaluation. When this choice task was first, only 37 percent of those who had undergone the training chose a female candidate. A similar pattern emerged when the order of the tasks was switched, in that participants were consistently biased on the first task and de-biased on the second, regardless of which task actually came first. One possible explanation for this effect is that participants seem to recognize that the researchers are trying to debias them, and then try to correct for this perceived influence by deliberately responding in more stereotypical ways, at least at first. Once they have an opportunity to explicitly counteract the debiasing, they stop trying to resist the training and then the effects emerge. Subsequently, they respond in counterstereotypical ways.

pothesized that interventions that encourage the perceiver to experience the emotions of the minority group might be effective de-biasing strategies. What does it feel like to have your intelligence automatically questioned, or to be trailed by detectives each time you walk into a store? Perspective-taking involves stepping into the shoes of a stereotyped person, and can be useful in assessing the emotional impact on individuals who are constantly being stereotyped in negative ways.

There are now multiple studies attesting to the merits of perspective taking as a strategy for reducing intergroup bias. Some have linked perspective-taking to decreased activation and application of negative group stereotypes (Galinsky and Moskowitz, 2000; Todd et al., 2011); others have shown that adopting the perspective of a particular out-group target leads to more positive evaluations of other individual members of the target's group (Shih et al., 2009) and of the target's group as a whole (Stephan and Finlay, 1999). For example, Todd et al. (2011) conducted a series of lab experiments examining the impact of perspective-taking on several outcomes: automatic evaluations, approach-avoidance reactions, and behaviors displayed during face-to-face interactions. In one of the experiments, participants watched a video depicting a series of discriminatory acts directed toward a black man versus a white man. As they watched the video, participants either adopted the black man's perspective or they attempted to remain objective and detached (control group). The researchers included two different perspective-taking conditions in this experiment. Some participants tried to imagine the black man's thoughts, feelings, and experiences (other condition) as they watched the video; others tried to imagine their own thoughts, feelings, and experiences as if they were in the black man's situation (self condition). After watching the video, participants completed a variant of the IAT that assesses automatic evaluations of black relative to white Americans. Subjects in both of the perspective-taking conditions (other and self conditions) exhibited significantly weaker pro-white bias than the control subjects.

Strategies targeting value-consistency and self-worth rely on the theory that individuals' desire to maintain consistency between valued cognitions and behaviors or protect their self-worth may be leveraged to lead them to repress their prejudice (Paluck and Green, 2009). De-biasing strategies in this area have leveraged cognitive dissonance and self-affirmation theories. For example, in a lab experiment, Leippe and Eisenstadt (1994) apply cognitive dissonance theory to get subjects to see prejudice as inconsistent with their own values: college students

softened their anti-black positions on social policies and reported more egalitarian attitudes and beliefs after agreeing to write a public statement in favor of pro-black policies. Also, Fein and Spencer (1997) report that lab subjects who have "self-affirmed" by circling values that were most important to them were more likely to give positive ratings to a Jewish job candidate.[44]

A body of research in social psychology suggests that prejudice and discrimination might also be influenced by expert opinion and greater accountability to others for one's beliefs and behaviors. Levy, Stroessner, and Dweck (1998) show that telling lab subjects that experts believe that personality is malleable reduces stereotyping against minority groups. Dobbs and Crano (2001) report that subjects allocated more points to a fictitious out-group when they were required to justify their allocations to others; similarly, Bodenhausen, Kramer, and Süsser (1994) show that students involved in a school disciplinary case were less likely to stereotype the student if they believed they would be accountable to their peers for their evaluation of the case.

Individuating is another socio-cognitive de-biasing strategy that involves gathering very specific information about a person's background, tastes, hobbies, and family, so that one's judgments will be based on the particulars of that person, rather than on group characteristics. This approach is grounded in the social identity and categorization literatures and essentially is a de-categorization effort, where subjects are instructed to focus on the individual rather than the group (Brewer, 1988; Fiske and Neuberg, 1990). Lebrecht et al. (2009) provide an interesting take on the individuation exercise. In their study, two groups of Caucasian subjects were exposed equally to the same African-American faces in a training protocol run over five sessions. In the individuation condition, subjects learned to discriminate between African-American faces; specifically, they received "expertise training" with other-race faces – defined by the authors as a procedure that improves observers' ability to individuate objects within the training domain and hence reduce the degree to which other-race faces are stereotyped. In contrast, in the categorization condition, subjects learned to categorize faces as African-American or not. Subjects in the individuation condition, but not in the categorization condition, showed improved discrimination of African-American faces with training. Also, subjects in the individuation condition, but not the categorization condition, showed a reduction in their implicit racial bias. For the individuation condition only, the degree to which an individual subject's implicit racial bias decreased was significantly correlated with the degree of improvement that the subject showed in their ability

[44]Note that participants who were Jewish were excluded from this part of the study.

to differentiate African-American faces.

Other de-biasing strategies inspired by the social identity and categorization literatures include re-categorization and crossed-categorization techniques, where participants are encouraged to think of people from different groups as part of one subordinate group using cues such as same shirt colors or shared prizes, or participants are made aware of their common membership in a third group. Such re-categorization and cross-categorization efforts have shown some success in reducing favoritism for the in-group and improving cooperation between groups (Dovidio and Gaertner, 2000; Gaertner et al., 1999).

An exciting recent study in the socio-cognitive de-biasing area is Lai et al. (2014), who sought to determine the effectiveness of various methods for reducing implicit bias. Structured as a research contest, teams of scholars were given five minutes in which to enact interventions that they believed would reduce implicit preferences for Whites compared to Blacks, as measured by an IAT, with the goal of attaining IAT scores that reflect a lack of implicit preference for either of the two groups. Teams submitted 18 interventions that were tested approximately two times across three studies, totaling 11,868 non-black participants. Half of the interventions were effective at reducing the implicit bias that favors Whites over Blacks. Most effective were the following interventions: (1) participating in a sports game in which all of the teammates were black while the opposing team was all-white and engaged in unfair play, and being subsequently instructed to recall how their black teammates helped them while their white opponents did not; (2) reading a graphic story in which one is asked to place oneself in the role of the victim who is assaulted by a white man and rescued by a black man; (3) practicing an IAT with counterstereotypic Black (e.g. Michael Jordan, Martin Luther King, Jr.) and counterstereotypic White (e.g., Timothy McVeigh, Jeffrey Dahmer) exemplars.

A concern one may have about the relevance of this lab evidence for the field is that it can only document fairly short-term effects (up to 24 hours), and hence might be of limited relevance to the real world. However, even such a short time frame might be relevant to some important decisions that have been shown to be subject to bias, such as human resource managers' decision on whether to pass on a given résumé, or teachers' grading decisions. Therefore, we believe that even short-term effects could be of real-world relevance.

What this lab evidence does not allow us to assess, however, is how these short-term impacts would differ if the same person (e.g. an HR manager) was repeatedly exposed to such de-biasing

strategies (e.g. every time he or she sits down to start reviewing résumés, or grading exams).

Some other de-biasing work in psychology has taken seriously this concern about one-shot, short-term interventions and has asked whether related strategies can be built to produce enduring reductions in bias. Work by Devine and a series of co-authors is of particular interest. Devine (1989) proposes a habit-breaking approach to prejudice reduction and likens implicit biases to deeply entrenched habits developed through socialization experiences. "Breaking the habit" of implicit bias therefore requires learning about the contexts that activate the bias and how to replace the biased responses with responses that reflect one's non-prejudiced goals. Devine and colleagues (Devine and Monteith, 1993; Plant and Devine, 2009) argue that the motivation to break the prejudice habit stems from two sources. First, people must be aware of their biases and they must also be concerned about the consequences of their biases before they will be motivated to exert effort to eliminate them. Second, people need to know when biased responses are likely to occur and how to replace those biased responses with ones more consistent with their goals.

Devine et al. (2012) develop and test a longer-term intervention to help people reduce implicit biases and "break the prejudice habit." The participants were 91 non-black introductory psychology students, who completed a 12-week longitudinal study for course credit. The key elements of the intervention were as follows. First, to ensure awareness of their bias, all participants completed a measure of implicit bias and received feedback about their level of bias. People assigned to the treatment group were also presented with a bias education and training program, the goals of which were to evoke a general concern about implicit biases and train people to eliminate them. The program lasted 45 minutes. The education component likened the expression of implicit biases to a habit and provided information linking implicit bias to discriminatory behaviors across a wide range of settings (e.g., interpersonal, employment, health). The training component described how to apply a variety of bias reduction strategies in daily life. The training section presented participants with a wide array of strategies (covering many of the strategies discussed below, such as taking the perspective of stigmatized others, imagining counter-stereotypic examples, training in negating stereotypical associations and individuation) as well as opportunities to engage in positive interactions with members of the minority group (e.g. inter-group contact). This enabled participants to flexibly choose the strategies most applicable to different situations in their lives.

Following the intervention, treated participants had lower IAT scores than control group par-

ticipants at 4 and 8 weeks after the intervention; moreover, the effects at 4 and 8 weeks were not systematically different from each other, indicating that the reduction in implicit race bias persisted over time. These data provide the first evidence that a controlled, randomized intervention can produce enduring reductions in implicit bias. The intervention created no changes in the participants' reported racial attitudes, but it did affect participants' concern about discrimination and their awareness of their personal bias. Also, concerns about discrimination emerged as a moderator for the interventions' effects. The intervention appears to have raised concerns about discrimination at week 2, and the biggest reduction in implicit bias in the treatment group was among those subjects who experienced growing concerns.

Despite the large amount of both theoretical and lab-based work in psychology on these various socio-cognitive de-biasing techniques, it is remarkable how few evaluations of these techniques have been performed in the field.

Paluck and Green (2009) perform a thorough literature search of the randomized field evidence on the de-biasing techniques listed above. While the number of field experiments they identify is non-trivial (71), much of the work they survey is not directly guided by the psychology literature or directory transposable into the specific lab-based strategies reviewed above. Moreover, very few of the existing field studies are designed to track changes in behavior outcome measures. The modal existing field study also involves a very short-term follow-up (often within the day) and takes place in a classroom setting with a student population, hence quite "lab-like" even if not explicitly in the lab.

By far most common have been interventions relying on various forms of entertainment (books, movies, cartoons, etc.) to create a persuasive narrative aimed at altering stereotypical thinking. In many cases though, the entertainment content is not based on the specific psychological theories that have guided the lab work, and it is hence difficult to make a direct link between the lab and the field evidence. For example, Paluck and Green (2009) identify several randomized field experiments performed in schools to measure the impact of reading on prejudice. This work suggests reduction in self-reported bias associated with reading content that portrays contact between children who are similar to the studied population and children of different race (e.g., intergroup friendship), as well as reading content that emphasizes a minority characters' individual characteristics rather than group membership (e.g., individuating). But is also possible that reading interventions might be effective because of the emotional reaction they

induce through perspective-taking (e.g., putting oneself in the shoes of the minority character in a book), or because they are a channel to communicate social norms (e.g., descriptions of what others are doing and hence what the reader should do).

Paluck and Green (2009) also identify a few instruction-based (rather than narrative-based) field interventions. In this case again, though, the content of the interventions is rarely directly guided by the lab evidence, and a lack of theoretical foundations may explain in part a lack of impressive findings. One exception is Lustig (2003), which evaluates a training program in Israel that aims to encourage perspective-taking and empathy to reduce prejudice against Palestinians among Jewish twelfth graders. Lustig (2003) reports encouraging findings among Jewish students who were asked to write an essay about the Israeli-Palestinian conflict from the Palestinian viewpoint.

The randomized field studies designed to directly test consciousness-raising, value consistency and self-worth, as well as re-, de- and cross-categorization techniques can essentially be counted on one hand.[45] All have been performed on student populations and have produced mixed results.

At the same time, we are confident that hundreds of anti-prejudice interventions directly inspired by the lab-based literature described above must be taking place yearly not only in schools but also in business and government settings, but are not being rigorously evaluated. It would be of first-order importance for researchers to strike up partnerships with organizations interested in better understanding the value of the diversity training programs they are investing resources in, both in terms of their immediate impact on bias and their ultimate impact on organizational performance. Human resource departments, police departments, and courtrooms are only a few of the possible real-world settings where a much-needed field validation of this large lab-based literature could be performed.

For example, the U.S. Department of Justice is funding the development of a curriculum for police staff that reflects on the Fair and Impartial Policing perspective. This training program applies the modern science of bias to policing: it trains officers on the effect of implicit bias and gives them the information and skills they need to reduce and manage their biases. The curriculum addresses not just racial/ethnic bias, but biases based on other factors such as gender,

[45] See Houlette et al. (2004) for re-categorization, Rokeach (1971) and Rokeach (1973) for value consistency, Katz and Zalk (1978) and Katz (2000) for cognitive retraining, and Lustig (2003) for perspective-taking.

sexual orientation, religion, socio-economic status, etc. Officers are taught skills, inspired by the lab-tested methods described above to reduce and manage their own biases. Social psychologists from around the nation who conduct the research on human biases are members of the team that help design the curriculum. While this program has been implemented with various target audiences (recruits/patrol officers, first line supervisors, mid-level managers, command staff and law enforcement trainers), to our knowledge it has not been the subject of a rigorous evaluation.

As another example, there has been much discussion in the recent years about how the socio-debiasing techniques described above could be used to de-bias judges and jurors. Kang et al. (2012) discuss possible ways to import these techniques to the courtroom. They argue that:

> In chambers and the courtroom buildings, photographs, posters, screen savers, pamphlets, and decorations ought to be used that bring to mind counter-typical exemplars or associations for participants in the trial process ... for jurors, then, ... the hope would be that by reminding them of counter-typical associations, we might momentarily activate different mental patterns while in the courthouse and reduce the impact of implicit biases on their decision-making.

Also, Elek and Agor (2014) show how de-biasing strategies could be feasibly brought to the courtroom with simple alterations to the standard instructions delivered by the judge to the jury, such as including a recognition of the universality of bias and explicit encouragement of perspective-taking.

3.4 Technological De-Biasing

Stanovich and West's (2000) distinction between System 1 and System 2 cognitive functioning provides a useful framework for organizing both what scholars have learned to date about effective strategies for improving decision-making and future efforts to uncover improvement strategies. System 1 refers to our intuitive system, which is typically fast, automatic, effortless, implicit, and emotional. System 2 refers to reasoning that is slower, conscious, effortful, explicit, and logical. People often lack important information regarding a decision, fail to notice available information, face time and cost constraints, and maintain a relatively small amount of information in their usable memory. The busier people are, the more they have on their minds, and the more time constraints they face, the more likely they will be to rely on System 1 thinking. In the many

situations where we know that decision biases are likely to plague us (e.g., when evaluating diverse job candidates, estimating our percent contribution to a group project, choosing between spending and saving, etc.), relying exclusively on System 1 thinking is likely to lead us to make errors. Also, when the basis for judgment is somewhat vague (e.g., situations that call for discretion, cases that involve the application of new, unfamiliar laws, etc.), biased judgments are more likely. Without more explicit, concrete criteria for decision-making, individuals tend to disambiguate the situation using whatever information is most easily accessible – including stereotypes (Dovidio and Gaertner, 2000; Johnson et al., 1995).

Similarly, certain emotional states (anger, disgust) can exacerbate implicit bias in judgments of stigmatized group members, even if the source of the negative emotion has nothing to do with the current situation or with the issue of social groups or stereotypes more broadly (DeSteno et al., 2004; Dasgupta et al., 2009). Interestingly, and perhaps counterintuitively, happiness may also produce more stereotypic judgments, though the exact reasoning for this is unclear and the stereotypic judgements can be consciously controlled if the person is motivated to do so (Bodenhausen, Kramer, and Süsser, 1994).

Circumstances that are tiring (e.g., long hours, fatigue), stressful (e.g., heavy, backlogged, or very diverse caseloads; loud construction noise; threats to physical safety; popular or political pressure about a particular decision; emergency or crisis situations), or otherwise distracting, can adversely affect judicial performance (Eells and Showalter, 1994; Hartley and Adams, 1974; Keinan, 1987). Specifically, situations that involve time pressure (Van Knippenberg et al., 1999), that force a decision-maker to form complex judgments relatively quickly (Bodenhausen and Lichtenstein, 1987), or in which the decision-maker is distracted and cannot fully attend to incoming information (Gilbert and Hixon, 1991; Sherman et al., 1998) all limit the ability to fully process case information. Decision-makers who are rushed, stressed, distracted, or pressured are more likely to apply stereotypes – recalling facts in ways biased by stereotypes and making more stereotypic judgments – than decision-makers whose cognitive abilities are not similarly constrained.

For instance, Correll et al. (2002) have used videogames in the lab to assess the effect of race on shoot/don't shoot decisions of targets that are either holding guns or holding non-threatening objects. While subjects are instructed to shoot the armed targets and not shoot the unarmed targets, subjects make errors and these errors are systematically correlated with the

race of the target: they disproportionately shoot unarmed Blacks and don't shoot armed Whites. Subsequent work has shown this "shooter bias" to be exacerbated when respondents are tired (Ma et al., 2013), rushed (Payne, 2006), or cannot see well (Payne, Shimizu, and Jacoby, 2005). Some of these circumstances are unavoidable during actual policing. However, any staffing and scheduling steps that minimize officer fatigue could also curb some of these racial disparities.

Danziger, Levav, and Avnaim-Pesso's (2011) field study of sequential parole decisions made by experienced judges provides another interesting illustration. Their sample is 1,112 parole board hearings in Israeli prisons, over a ten-month period. The rulings were made by eight Jewish-Israeli judges, with an average of 22 years of judging behind them. Every day, each judge considers between 14 and 35 cases, spending around six minutes on each decision. They take two food breaks that divide their day into three sessions. All of these details, from the decision to the times of the breaks, are duly recorded. They record the judges' two daily food breaks, which result in segmenting the deliberations of the day into three distinct "decision sessions." They find that the percentage of favorable rulings drops gradually from 65 percent to nearly zero within each decision session and returns abruptly to 65 percent after a break. The researchers attribute their results to the repetitive decision-making tasks draining the judges' mental resources. When drained, the judges start suffering from "choice overload" and start opting for the easiest default choice, which is to deny the prisoner's request. The more decisions a judge has made, the more drained he or she is, and the more likely the judge will make the default choice. Taking a break replenishes him or her. However, the researchers did not find any evidence that the timing of the decision affected discrimination: judges treated the prisoners equally regardless of their gender and ethnicity, as well as the severity of their crime.

Casey et al. (2012) study how one could build on this knowledge of what triggers System 1 versus System 2 thinking to help technologically de-bias the courtroom. For example, the authors discuss how jurors might be allowed more time on cases in which implicit bias might be a concern by, for example, spending more time reviewing the facts of the case before committing to a decision; similarly, courts may review areas in which judges and other decision-makers are likely to be over-burdened and consider options (e.g., reorganizing court calendars) for modifying procedures to provide more time for decision-making. Also, jurors may be asked to commit to decision-making criteria before reviewing case-specific information and courts may develop protocols that identify potential sources of ambiguity. Furthermore, courtrooms could consider the

pros (e.g., more understanding of issues) and cons (e.g., familiarity may lead to less deliberative processing) of using judges with special expertise to handle cases with greater ambiguity. A lot of the possible strategies that Casey et al. (2012) discuss for the courtroom setting could naturally be applied to other real-world setting where biases in decision-making have been documented.

Another strategy for moving toward System 2 thinking might be, in settings where data exists on past inputs to and outcomes from a particular decision-making process, to have decision-makers construct a linear model, or a formula that weights and sums the relevant predictor variables to reach a quantitative forecast about the outcome. Researchers have found that linear models produce predictions that are superior to those of experts across an impressive array of domains (Dawes, 1971).[46] In general, the use of linear or more complex machine learning models can help decision-makers avoid the pitfalls of many judgment biases, yet this method has only been tested in a small subset of the potentially relevant domains.

With better knowledge of why discrimination occurs in a particular setting, it will become easier to design appropriate technological de-biasing strategies. As we discussed earlier in Section 1.5, Bartoš et al. (2013) convincingly demonstrate racial gaps in attention allocation by HR managers. Once they see a minority name on a résumé, they pay less attention to that résumé. These findings confirm the merit of requiring separate rankings of applicants from non-minority and minority groups (or across gender lines) followed by a comparison of leading candidates across the groups. One can think of this rule as providing quotas in the pre-selection process. We do not know of any systematic evaluation of such a strategy.

Also, since the earlier a decision-maker learns a group attribute, such as name, the larger the asymmetry in attention to subsequent information such as education or qualification, the findings in Bartoš et al. (2013) strengthens the case for suppressing the signals of a group attribute during the part of the selection process. This particular technological approach has been receiving quite a lot of attention from policymakers in the recent years and has been evaluated in the field. In particular, the large number of correspondence studies have raised interest in the possibility of using "blind" hiring procedures. In some recruiting circumstances, the full hiring process can take place anonymously. Goldin and Rouse (2000) famously showed that American orchestras conducting blind auditions hired more women. In most other cases, though, only the first stage

[46]The value of linear models in hiring, admissions, and selection decisions is highlighted by research that Moore et al. (2010) conducted on the interpretation of grades by graduate admission officers.

of the recruitment is made anonymous: this is the case in anonymous application procedures, such as the masking of identifying characteristics in résumés at the first selection stage.

In several European countries, pilot studies of the impact of such anonymization of résumés have been conducted, including relatively large-scale field experiments in France, the Netherlands, Sweden and Germany. These experiments are summarized in Krause, Rinne, and Zimmermann (2012b). Only a subset were truly randomized and we focus our discussion on this subset.[47]

In 2010 and 2011, the French government initiated an experiment, which was implemented by the French public employment service. It involved about 1,000 firms in eight local labor markets and it lasted in total for about ten months (Behaghel, Crépon, and Le Barbanchon, 2014).

Among volunteer firms, résumés were either transmitted anonymously or non-anonymously. The experiment's main findings can be summarized as follows. First, women benefit from higher call-back rates with anonymous job applications – at least if they compete with male applicants for a job. Second, and most interestingly, migrants and residents of deprived neighborhoods suffer from anonymous job applications. Their call-back rates are lower with anonymous job applications than with standard applications. This adverse effect on minority candidates is the exact opposite effect to what policymakers had hoped, and a surprising result given existing evidence from correspondence testing in France (Duguet et al., 2010), which shows discrimination against minority candidates for some jobs, no discrimination for others, but never discrimination against majority candidates. Behaghel, Crépon, and Le Barbanchon (2014) explain these surprising results by the self-selection of firms that agreed to participate in the field experiment. Among firms that were contacted to participate in the experiment, 62 percent accepted the invitation. While participating firms were very similar to refusing firms in most observable dimensions, there was one significant exception: participating firms tended to interview and hire relatively more minority candidates (when using standard résumés). The anonymization therefore prevented selected firms from treating minority candidates more favorably during the experiment. Hence,

[47] Åslund and Skans (2012) analyze an experiment conducted in parts of the local administration of the Swedish city of Gothenburg between 2004 and 2006. Based on a difference-in-differences approach, the authors find that anonymous job applications increase the chances of an interview invitation for both women and applicants of non-Western origin when compared to standard applications. These increased chances for minority candidates in the first stage also translated into a higher job offer arrival rate for women, but not for migrants. In the Netherlands, two experiments took place in the public administration of one major Dutch city in 2006 and 2007 (Bøg and Kranendonk, 2011). The experiments focused on ethnic minorities. The lower call-back rate for minority candidates with standard applications disappears with anonymous job applications. With regards to job offers, however, the authors do not detect any differences between minority and majority candidates – irrespective of whether or not their résumés are treated anonymously.

the results of the experiment cannot be viewed as representative of what anonymization might have achieved if it had been mandated to all firms. Methodologically, this paper offers a valuable illustration of one danger when trying to generalize the findings of a field experiment. External validity is far from guaranteed if there is sizable room for selection or self-selection of subjects into the experiment (Heckman, 1992; Allcott, 2015).

Another large-scale randomized field experiment took place in Germany in early 2010 (Krause, Rinne, and Zimmermann, 2012a). The publication of a correspondence testing study for Germany (Kaas and Manger, 2012) triggered a lively public debate about discrimination in the hiring decisions of German firms.[48] Against this background, the Federal Anti-Discrimination Agency initiated a field experiment with anonymous job applications in Germany to investigate their potential in combating hiring discrimination. This experiment was also subject to selection in participation, with eight organizations voluntarily joining the experiment. The characteristics that were made anonymous include the applicant's name and contact details, gender, nationality, date and place of birth, disability, marital status and the applicant's picture.[49] Unlike the French study, the authors find that the anonymization leads to less discrimination against minority groups. Moreover, anonymizing applications is not too difficult administratively, with standardized application forms that are completed by the applicants appearing as the most effective and efficient way to make applications anonymous.

Conclusion

We have organized this chapter along three overarching themes: the measurement of discrimination, the consequences of discrimination, and factors and policies that may help undermine it. It is apparent from our review of the existing field experiments under each of these themes that there remain more unanswered or unexplored questions than there are settled ones.

By far the bulk of the field experiments that have been conducted in this area relate to the

[48] The study finds that applicants with a Turkish-sounding name are on average 14 percentage points less likely to receive an invitation for a job interview than applicants with a German-sounding name who are otherwise similar. In small- and medium-sized firms, this difference is even larger and amounts to 24 percentage points.

[49] The study was further designed to assess the practicality of different methods to remove identifiers from applications; practicality was assessed from interviews with the HR specialists at the firms. Four methods were considered: (1) standardized application forms in which sensitive information is not included; (2) refinements of existing online application forms such that sensitive information is disabled; (3) copying applicant's non-sensitive information into another document; (4) blacking out sensitive information in the original application documents.

measurement of discrimination using the correspondence method. This body of work has demonstrated how remarkably pervasive the differential treatment of minority groups is throughout the world (at least in the labor market and rental market). These studies, most often focusing on a single minority group in a single country, have been important in generating debates in the local media and local public opinion and, from that perspective, each has added value. In many fields of inquiry, researchers shy away from replication, but this is refreshingly not the case here – most likely because demonstrating differential treatment in the given country seemed a sufficiently important goal. On the other hand, researchers' ability to push the correspondence method further to go beyond pure measurement of differential treatment has been more limited. Disappointingly, there has been minimal methodological innovation in the way correspondence studies are being carried out. The main innovation might have been in leveraging the method to study differential treatment across other characteristics than race, gender or ethnicity, such as in the set of recent studies using the method to study discrimination against the long-term unemployed. While one might conclude from this that the correspondence method might have reached its full potential, recent papers such as Bartoš et al. (2013) which demonstrate how it can used to study the dynamics of discrimination (endogenous attention allocation in this case) suggest there remain unexplored avenues for more creative uses.

Perhaps because so much of economists' attention has been devoted to using field experiments to measure the extent of discrimination, there has been much less activity in designing creative ways to better document either its consequences or ways to undermine it. The dearth of field-based evidence on these last two themes is particularly striking given the rich theoretical and lab-based literatures (mainly in psychology) that such work could build upon. On the topic of consequence of discrimination, we are heartened to see a few recent papers such as Glover, Pallais, and Pariente (2015) that develop a creative field design to demonstrate how discrimination can be self-perpetuating. We believe that the last theme in our chapter, interventions to undermine discrimination, is particularly ripe for more field experimentation. It is striking that most of the research in economics on this question has centered around the contact hypothesis and exposure effects, while so many other strategies to de-bias people have been proposed by psychologists and evaluated in the lab. We strongly encourage researchers to take on this work in the near future. Creating more partnerships with organizations that are willing to provide the testing ground for different de-biasing strategies will be particularly useful for this work to move forward. More

generally, while field experiments in the last decade have been instrumental in documenting the prevalence of discrimination, field experiments in the future decade should aim to play as large of a role in isolating effective methods to combat it.

References

Jason Abrevaya and Daniel S. Hamermesh. Charity and Favoritism in the Field: Are Female Economists Nicer (to Each Other)? *Review of Economic and Statistics*, 94(1):202–207, 2012.

Alessandro Acquisti and Christina M. Fong. An Experiment in Hiring Discrimination via Online Social Networks. *Available at SSRN: http://ssrn.com/abstract=2031979 or http://dx.doi.org/10.2139/ssrn.2031979*, 2013.

Kenneth R. Ahern and Amy K. Dittmar. The Changing of the Boards: The Impact on Firm Valuation of Mandated Female Board Representation. *The Quarterly Journal of Economics*, 127(1):137–197, 2012.

Ali M. Ahmed and Mats Hammarstedt. Discrimination in the Rental Housing Market: A Field Experiment on the Internet. *Journal of Urban Economics*, 64(2):362–372, 2008.

Ali M. Ahmed and Mats Hammarstedt. Detecting Discrimination against Homosexuals: Evidence from a Field Experiment on the Internet. *Economica*, 76(303):588–597, 2009.

Ali M. Ahmed, Lina Andersson, and Mats Hammarstedt. Can Discrimination in the Housing Market Be Reduced by Increasing the Information about the Applicants? *Land Economics*, 86(1):79–90, 2010.

Ali M. Ahmed, Lina Andersson, and Mats Hammarstedt. Does Age Matter for Employability? A Field Experiment on Ageism in the Swedish Labour Market. *Applied Economics Letters*, 19 (4):403–406, 2012.

Ali M. Ahmed, Lina Andersson, and Mats Hammarstedt. Are Gay Men and Lesbians Discriminated against in the Hiring Process? *Southern Economic Journal*, 79(3):565–585, 2013.

Dennis J. Aigner and Glen G. Cain. Statistical Theories of Discrimination in Labor Markets. *Industrial and Labor Relations Review*, pages 175–187, 1977.

Alberto Alesina and Eliana La Ferrara. Ethnic Diversity and Economic Performance. *Journal of Economic Literature*, 43:762–800, 2005.

Hunt Allcott. Site Selection Bias in Program Evaluation. *The Quarterly Journal of Economics*, 130(3), 2015.

Gordon W. Allport. *The Nature of Prejudice*. Addison-Wesley, Cambridge, MA, 1954.

Joseph G. Altonji and Rebecca M. Blank. Chapter 48: Race and Gender in the Labor Market. *Handbook of Labor Economics*, 3(C):3143–3259, 1999.

Hal R. Arkes and Philip E. Tetlock. Attributions of Implicit Prejudice, or "Would Jesse Jackson 'Fail' the Implicit Association Test?". *Psychological Inquiry*, 15(4):257–278, 2004.

Joshua Aronson, Carrie B. Fried, and Catherine Good. Reducing the Effects of Stereotype Threat on African American College Students by Shaping Theories of Intelligence. *Journal of Experimental Social Psychology*, 38(2):113–125, 2002.

Kenneth J. Arrow. The Theory of Discrimination. *Discrimination in Labor Markets*, 3(10):3–33, 1973.

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

Ian Ayres and Peter Siegelman. Race and Gender Discrimination in Bargaining for a New Car. *American Economic Review*, 85(3):304–321, 1995.

Elisha Y Babad, Jacinto Inbar, and Robert Rosenthal. Pygmalion, Galatea, and the Golem: Investigations of Biased and Unbiased Teachers. *Journal of Educational Psychology*, 74(4): 459, 1982.

Stijn Baert, Bart Cockx, Niels Gheyle, and Cora Vandamme. Do Employers Discriminate Less If Vacancies Are Difficult to Fill? Evidence from a Field Experiment. *IZA Discussion Paper*, pages 1–30, 2013.

Manuel Bagues and Maria J. Perez-Villadoniga. Do Recruiters Prefer Applicants with Similar Skills? Evidence from a Randomized Natural Experiment. *Journal of Economic Behavior & Organization*, 82(1):12–20, 2012.

Manuel Bagues and Natalie Zinovyeva. The Role of Connections in Academic Promotions. *American Economic Journal: Applied Economics*, 7(2):264–292, 2015.

Manuel Bagues, Mauro Sylos-Labini, and Natalia Zinovyeva. Do Gender Quotas Pass the Test? Evidence from Academic Evaluations in Italy. *Scuola Superiore Sant'Anna, LEM Working Paper Series*, 14, 2014.

Manuel F. Bagues and Berta Esteve-Volart. Can Gender Parity Break the Glass Ceiling? Evidence from a Repeated Randomized Experiment. *Review of Economic Studies*, 77(4):1301–1328, 2010.

John Bailey, Michael Wallace, and Bradley Wright. Are Gay Men and Lesbians Discriminated Against When Applying for Jobs? A Four-City, Internet-Based Field Experiment. *Journal of Homosexuality*, 60(6):873–894, 2013.

Massimo Baldini and Marta Federici. Ethnic Discrimination in the Italian Rental Housing Market. *Journal of Housing Economics*, 20(1):1–14, 2011.

Mahzarin Banaji, Brian A. Nosek, and Anthony G. Greenwald. No Place for Nostalgia in Science: A Response to Arkes and Tetlock. *Psychological Inquiry*, 15(4):279–310, 2004.

Mahzarin R. Banaji and Anthony G. Greenwald. Implicit Gender Stereotyping in Judgments of Fame. *Journal of Personality and Social Psychology*, 68(2):181, 1995.

Abhijit Banerjee and Rohini Pande. Parochial Politics: Ethnic Preferences and Politician Corruption. *CEPR Discussion Paper No. DP6381*, 2009.

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

Abhijit Banerjee, Donald Green, Jennifer Green, and Rohini Pande. Can Voters be Primed to Choose Better Legislators? Experimental Evidence from Rural India. Unpublished manuscript, available at http://www.povertyactionlab.org/node/2764, 2010.

Abhijit Banerjee, Esther Duflo, Clement Imbert, and Rohini Pande. Entry, Exit, and Candidate Selection: Evidence from India. *Mimeo, 3ie Grantee Final Report*, 2013.

Pranab K. Bardhan, Dilip Mookherjee, and Monica Parra Torrado. Impact of Political Reservations in West Bengal Local Governments on Anti-Poverty Targeting. *Journal of Globalization and Development*, 1(1), 2010.

Vojtěch Bartoš, Michal Bauer, Julie Chytilová, and Filip Matějka. Attention Discrimination: Theory and Field Experiments. *CERGE Working Paper, ISSN 1211-3298*, 2013.

Lori Beaman, Raghabendra Chattopadhyay, Esther Duflo, Rohini Pande, and Petia Topalova. Powerful Women: Does Exposure Reduce Bias? *The Quarterly Journal of Economics*, 124(4): 1497–1540, 2009.

Lori Beaman, Esther Duflo, Rohini Pande, and Petia Topalova. Political Reservation and Substantive Representation: Evidence from Indian Village Councils. In Suman Berry, Barry Bosworth, and Arvind Panagariya, editors, *India Policy Forum 2010-11, Volume 7*. SAGE Publications Inc., 2010.

Lori Beaman, Esther Duflo, Rohini Pande, and Petia Topalova. Female Leadership Raises Aspirations and Educational Attainment for Girls: A Policy Experiment in India. *Science*, 335 (6068):582–586, 2012.

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

Luc Behaghel, Bruno Crépon, and Thomas Le Barbanchon. Unintended Effects of Anonymous Resumes. *CEPR Discussion Paper No. DP10215*, 2014.

Daniel J. Benjamin, James J. Choi, and A. Joshua Strickland. Social Identity and Preferences. *American Economic Review*, 100:1913–1928, 2010.

Marianne Bertrand and Sendhil Mullainathan. Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review*, 94(4):991–1013, 2004.

Marianne Bertrand and Antoinette Schoar. Managing with Style: the Effect of Managers on Firm Policies. *Quarterly Journal of Economics*, 118(4):1169–208, 2003.

Marianne Bertrand, Dolly Chugh, and Sendhil Mullainathan. Implicit Discrimination. *American Economic Review*, pages 94–98, 2005.

Timothy Besley and Stephen Coate. An Economic Model of Representative Democracy. *The Quarterly Journal of Economics*, pages 85–114, 1997.

Timothy J. Besley, Olle Folke, Torsten Persson, and Johanna Rickne. Gender Quotas and the Crisis of the Mediocre Man: Theory and Evidence from Sweden. *IFN Working Paper*, 2013.

Diana E. Betz and Denise Sekaquaptewa. My Fair Physicist? Feminine Math and Science Role Models Demotivate Young Girls. *Social Psychological and Personality Science*, 3(6):738–746, 2012.

Rikhil R. Bhavnani. Do Electoral Quotas Work after They Are Withdrawn? Evidence from a Natural Experiment in India. *American Political Science Review*, 103(01):23, 2009.

Jeff E. Biddle and Daniel S. Hamermesh. Beauty, Productivity and Discrimination: Laywers' Looks and Lucre. *Journal of Labor Economics*, 15:172–201, 1998.

Lisa S. Blackwell, Kali H. Trzesniewski, and Carol Sorich Dweck. Implicit Theories of Intelligience Predict Achievement Across an Adolescent Transition: A Longitudinal Study and an Intervention. *Child Development*, 78(1):246–263, 2007.

Lieselotte Blommaert, Marcel Coenders, and Frank van Tubergen. Discrimination of Arabic-Named Applicants in the Netherlands: An Internet-Based Field Experiment Examining Different Phases in Online Recruitment Procedures. *Social Forces*, 92(3):957–982, 2014.

Galen V. Bodenhausen and Meryl Lichtenstein. Social Stereotypes and Information-Processing Strategies: The Impact of Task Complexity. *Journal of Personality and Social Psychology*, 52 (5):871, 1987.

Galen V. Bodenhausen, Geoffrey P. Kramer, and Karin Süsser. Happiness and Stereotypic Thinking in Social Judgment. *Journal of Personality and Social Psychology*, 66(4):621, 1994.

Martin Bøg and Erik Kranendonk. Labor Market Discrimination of Minorities? Yes, but not in Job Offers. *MPRA Paper No. 33332*, 2011.

Johanne Boisjoly, Greg J. Duncan, Michael Kremer, Dan M. Levy, and Jacque Eccles. Empathy or Antipathy? The Impact of Diversity. *American Economic Review*, 96(5):1890–1905, 2006.

Alison Booth and Andrew Leigh. Do Employers Discriminate by Gender? A Field Experiment in Female-Dominated Occupations. *Economics Letters*, 107(2):236–238, 2010.

Alison L. Booth, Andrew Leigh, and Elena Varganova. Does Ethnic Discrimination Vary Across Minority Groups? Evidence from a Field Experiment. *Oxford Bulletin of Economics and Statistics*, 74(4):547–573, 2011.

Mariano Bosch, M. Angeles Carnero, and Lídia Farré. Information and Discrimination in the Rental Housing Market: Evidence from a Field Experiment. *Regional Science and Urban Economics*, 40(1):11–19, 2010.

Anne Boschini, Astri Muren, and Mats Persson. Constructing Gender in the Economics Lab. Technical report, Stockholm University, Department of Economics, 2009.

Marilynn B. Brewer. Ethnocentrism and Its Role in Interpersonal Trust. *Scientific Inquiry and the Social Sciences*, 214, 1981.

Marilynn B. Brewer. A Dual Process Model of Impression Formation. In R. Wyer and T. Srull, editors, *Advances in Social Cognition*, volume 1, pages 1–36. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, 1988.

Ivy E. Broder. Review of NSF Economics Proposals: Gender and Institutional Patterns. *The American Economic Review*, pages 964–970, 1993.

Colin Brown and Pat Gay. *Racial Discrimination: 17 Years after the Act*. Policy Studies Institute, 1985.

Justine Burns, Lucia Corno, and Eliana La Ferrara. Interaction, Prejudice and Performance. Evidence from South Africa. *Working Paper*, 2015.

Moa Bursell. What's in a Name? A Field Experiment Test for the Existence of Ethnic Discrimination in the Hiring Process. *SULCIS WP*, 7, 2007.

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

Magnus Carlsson and Stefan Eriksson. Discrimination in the Rental Market for Apartments. *Journal of Housing Economics*, 23:41–54, 2014.

Adrian G. Carpusor and William E. Loges. Rental Discrimination and Ethnicity in Names. *Journal of Applied Social Psychology*, 36(4):934–952, 2006.

Pamela M. Casey, Roger K. Warren, Fred L. Cheesman, and Jennifer K. Elek. Helping Courts Address Implicit Bias. Technical report, National Center for State Courts, Williamsburg, VA, 2012.

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

Raghabendra Chattopadhyay and Esther Duflo. Women as Policy Makers: Evidence from a Randomized Experiment in India. *Econometrica*, 72:1409–1443, 2004.

Sapna Cheryan, John Oliver Siy, Marissa Vichayapai, Benjamin J. Drury, and Saenam Kim. Do Female and Male Role Models who Embody STEM Stereotypes Hinder Women's Anticipated Success in STEM? *Social Psychological and Personality Science*, 2(6):656–664, 2011.

Irma Clots-Figueras. Are Female Leaders Good for Education? Evidence from India. *Mimeo, Universidad Carlos III de Madrid*, 2009.

Irma Clots-Figueras. Women in Politics: Evidence from the Indian States. *Journal of Public Economics*, 95(7-8):664–690, 2011.

Stephen Coate and Glenn Loury. Antidiscrimination Enforcement and the Problem of Patronization. *The American Economic Review*, pages 92–98, 1993.

Katherine B. Coffman, Lucas C. Coffman, and Keith M. Marzilli Ericson. The Size of the LGBT Population and the Magnitude of Anti-Gay Sentiment are Substantially Underestimated. *NBER Working Paper No. 19508*, 2013.

Geoffrey L. Cohen, Julio Garcia, Nancy Apfel, and Allison Master. Reducing the Racial Achievement Gap: A Social-Psychological Intervention. *Science*, 313(5791):1307–1310, 2006.

Geoffrey L. Cohen, Julio Garcia, Valerie Purdie-Vaughns, Nancy Apfel, and Patricia Brzustoski. Recursive Processes in Self-Affirmation: Intervening to Close the Minority Achievement Gap. *Science*, 324(5925):400–403, 2009.

Joshua Correll, Bernadette Park, Charles M. Judd, and Bernd Wittenbrink. The Police Officer's Dilemma: Using Ethnicity to Disambiguate Potentially Threatening Individuals. *Journal of Personality and Social Psychology*, 83(6):1314–1329, 2002.

Harry Cross, Genevieve Kenney, Jane Mell, and Wendy Zimmerman. Employer Hiring Practices: Differential Treatment of Hispanic and Anglo Job Seekers. *Urban Institute Report 90-4*, 1990.

Shai Danziger, Jonathan Levav, and Liora Avnaim-Pesso. Extraneous Factors in Judicial Decisions. *Proceedings of the National Academy of Sciences*, 108(17):6889–6892, 2011.

Nilanjana Dasgupta and Shaki Asgari. Seeing is Believing: Exposure to Counterstereotypic Women Leaders and Its Effect on the Malleability of Automatic Gender Stereotyping. *Journal of Experimental Social Psychology*, 40(5):642–658, 2004.

Nilanjana Dasgupta and Anthony G. Greenwald. On the Malleability of Automatic Attitudes: Combating Automatic Prejudice with Images of Admired and Disliked Individuals. *Journal of Personality and Social Psychology*, 81(5):800, 2001.

Nilanjana Dasgupta, David DeSteno, Lisa A. Williams, and Matthew Hunsinger. Fanning the Flames of Prejudice: The Influence of Specific Incidental Emotions on Implicit Prejudice. *Emotion*, 9(4):585, 2009.

Robyn M. Dawes. A Case Study of Graduate Admissions: Application of Three Principles of Human Decision Making. *American Psychologist*, 26:180–188, 1971.

Thomas S. Dee. Stereotype Threat and the Student-Athlete. *NBER Working Paper No. 14705*, 2009.

David DeSteno, Nilanjana Dasgupta, Monica Y. Bartlett, and Aida Cajdric. Prejudice From Thin Air: The Effect of Emotion on Automatic Intergroup Attitudes. *Psychological Science*, 15(5):319–324, 2004.

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

Patricia G. Devine and Margo J. Monteith. The Role of Discrepancy-Associated Affect in Prejudice Reduction. In Diane M. Mackie and David L. Hamilton, editors, *Affect, Cognition, and*

Stereotyping: Interactive Processes in Group Perception, pages 317–344. Academic Press, San Diego, CA, 1993.

Patricia G. Devine, Patrick S. Forscher, Anthony J. Austin, and William T.L. Cox. Long-term Reduction in Implicit Race Bias: A Prejudice Habit-Breaking Intervention. *Journal of Experimental Social Psychology*, 48(6):1267–1278, 2012.

Eric S. Dickson and Kenneth Scheve. Social Identity, Political Speech, and Electoral Competition. *Journal of Theoretical Politics*, 18(1):5–39, 2006.

John Dixon, Mark Levine, Steve Reicher, and Kevin Durrheim. Beyond Prejudice: Are Negative Evaluations the Problem and is Getting Us to Like One Another More the Solution? *Behavioral and Brain Sciences*, 35(6):411–425, 2012.

Michael Dobbs and William D. Crano. Outgroup Accountability in the Minimal Group Paradigm: Implications for Aversive Discrimination and Social Identity Theory. *Journal of Personality and Social Psychology*, 27:355–364, 2001.

Jennifer L. Doleac and Luke C.D. Stein. The Visible Hand: Race and Online Market Outcomes. *The Economic Journal*, 123(572):F469–F492, 2013.

John F. Dovidio. On the Nature of Contemporary Prejudice: The Third Wave. *Journal of Social Issues*, 57(4):829–849, 2001.

John F. Dovidio and Samuel L. Gaertner. Aversive Racism and Selection Decisions: 1989 and 1999. *Psychological Science*, 11(4):315–319, 2000.

John F. Dovidio, Samuel L. Gaertner, Alice M. Isen, Mary Rust, and Paula Guerra. Positive Affect, Cognition, and the Reduction of Intergroup Bias. In Constantine Sedikides, John Schopler, and Chester A. Insko, editors, *Intergroup Cognition and Intergroup Behavior*, pages 337–366. Lawrence Erlbaum Associates, Mahwah, NJ, 1998a.

John F. Dovidio, Samuel L. Gaertner, and Ana Validzic. Intergroup Bias: Status, Differentiation, and a Common In-Group Identity. *Journal of Personality and Social Psychology*, 75(1):109, 1998b.

John F. Dovidio, Samuel L. Gaertner, and Tamar Saguy. Commonality and the Complexity of "We": Social Attitudes and Social Change. *Personality and Social Psychology Review*, 13(1): 3–20, 2009.

Judith Droitcour, Rachel A. Caspar, Michael L. Hubbard, and Trena M. Ezzati. The Item Count Technique as a Method of Indirect Questioning: A Review of Its Development and a Case Study Application. In P.B. Beimer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz, and S. Sudman, editors, *Measurement Errors in Surveys*, pages 185–211. John Wiley & Sons, Inc, Hoboken, NJ, 1991.

Emmanuel Duguet, Noam Leandri, Yannick L'horty, and Pascale Petit. Are Young French Jobseekers of Ethnic Immigrant Origin Discriminated Against? A Controlled Experiment in the Paris Area. *Annals of Economics and Statistics*, pages 187–215, 2010.

Thad Dunning and Janhavi Nilekani. Ethnic Quotas and Political Mobilization: Caste, Parties, and Distribution in Indian Village Councils. *American Political Science Review*, 107(01):35–56, 2013.

J.B. Dusek, V.C. Hall, and W.J. Meyer. *Teacher Expectations*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1985.

Alice H. Eagly and Steven J. Karau. Role Congruity Theory of Prejudice toward Female Leaders. *Psychological Review*, 109(3):573, 2002.

Alice H. Eagly, Mona G. Makhijani, and Bruce G. Klonsky. Gender and the Evaluation of Leaders: A Meta-Analysis. *Psychological Bulletin*, 111(1):3, 1992.

Dov Eden. Leadership and Expectations: Pygmalion Effects and Other Self-Fulfilling Prophecies in Organizations. *The Leadership Quarterly*, 3(4):271–305, 1992.

Dov Eden and Gad Ravid. Pygmalion vs. Self-Expectancy: Effects of Instructor- and Self-Expectancy on Trainee Performance. *Organization Behavior and Human Performance*, 30: 351–364, 1982.

Dov Eden and Abraham B. Shani. Pygmalion Goes to Boot Camp: Expectancy, Leadership, and Trainee Performance. *Journal of Applied Psychology*, 67(2):194, 1982.

Tracy D. Eells and C. Robert Showalter. Work-Related Stress in American Trial Judges. *Journal of the American Academy of Psychiatry and the Law Online*, 22(1):71–83, 1994.

Jennifer K. Elek and Paula Hannaford Agor. Can Explicit Instructions Reduce Expression of Implicit Bias? New Questions Following a Test of a Specialized Jury Instruction. Available at SSRN: http://ssrn.com/abstract=2430438, April 28 2014.

David Engel, Anita Williams Woolley, Lisa X. Jing, Christopher F. Chabris, and Thomas W. Malone. Reading the Mind in the Eyes or Reading between the Lines? Theory of Mind Predicts Collective Intelligence Equally Well Online and Face-To-Face. *PLoS one*, 9(12):e115212, 2014.

Stefan Eriksson and Dan-Olof Rooth. Do Employers Use Unemployment as a Sorting Criterion When Hiring? Evidence from a Field Experiment. *American Economic Review*, 104(3):1014–1039, 2014.

Michael Ewens, Bryan Tomlin, and Liang Choon Wang. Statistical Discrimination or Prejudice? A Large Sample Field Experiment. *Review of Economics and Statistics*, 96(1):119–134, 2014.

Steven Fein and Steven J. Spencer. Prejudice as Self-Image Maintenance: Affirming the Self through Derogating Others. *Journal of Personality and Social Psychology*, 73(1):31, 1997.

Susan T. Fiske and Steven L. Neuberg. A Continuum of Impression Formation from Category-Based to Individuating Processes: Influences of Information and Motivation on Attention and Interpretation. *Advances in Experimental Social Psychology*, 23:1–74, 1990.

Raymond Fisman, Sheena S. Iyengar, Emir Kamenica, and Itamar Simonson. Racial Preferences in Dating. *Review of Economic Studies*, 75(177-32), 2008.

Michael Fix and Raymond J. Struyk. *Clear and Convincing Evidence: Measurement of Discrimination in America*. Urban Institute Press, 1993.

Chad E. Forbes and Toni Schmader. Retraining Attitudes and Stereotypes to Affect Motivation and Cognitive Capacity under Stereotype Threat. *Journal of Personality and Social Psychology*, 99(5):740, 2010.

Mark R. Forehand, Rohit Deshpandé, and Americus Reed II. Identity Salience and the Influence of Differential Activation of the Social Self-Schema on Advertising Response. *Journal of Applied Psychology*, 87(6):1086–1099, 2002.

Roland G. Fryer and Steven D. Levitt. The Causes and Consequences of Distinctively Black Names. *The Quarterly Journal of Economics*, 119(3):767–805, 2004.

Samuel L. Gaertner, John F. Dovidio, Mary C. Rust, Jason A. Nier, Brenda S. Banker, Christine M. Ward, Gary R. Mottola, and Missy Houlette. Reducing Intergroup Bias: Elements of Intergroup Cooperation. *Journal of Personality and Social Psychology*, 76(3):388, 1999.

Francisco B. Galarza and Gustavo Yamada. Labor Market Discrimination in Lima, Peru: Evidence from a Field Experiment. *World Development*, 58:83–94, 2014.

Adam D. Galinsky and Gordon B. Moskowitz. Perspective-Taking: Decreasing Stereotype Expression, Stereotype Accessibility, and In-Group Favoritism. *Journal of Personality and Social Psychology*, 78(4):208, 2000.

George Galster. Racial Discrimination in Housing Markets during the 1980s: A Review of the Audit Evidence. *Journal of Planning Education and Research*, 9(3):165–175, 1990.

Bertram Gawronski, Roland Deutsch, Sawsan Mbirkou, Beate Seibt, and Fritz Strack. When "Just Say No" is Not Enough: Affirmation versus Negation Training and the Reduction of Automatic Stereotype Activation. *Journal of Experimental Social Psychology*, 44(2):370–377, 2008.

Rand Ghayad. The Jobless Trap. *Job Market Paper*, pages 1–39, 2013.

Daniel T. Gilbert and J. Gregory Hixon. The Trouble of Thinking: Activation and Application of Stereotypic Beliefs. *Journal of Personality and Social Psychology*, 60(4):509, 1991.

Demis E. Glasford and Justine Calcagno. The Conflict of Harmony: Intergroup Contact, Commonality and Political Solidarity between Minority Groups. *Journal of Experimental Social Psychology*, 48(1):323–328, 2012.

Dylan Glover, Amanda Pallais, and William Pariente. Discrimination as a Self-Fulfilling Prophecy: Evidence from French Grocery Store. *Working Paper*, 2015.

Philip Goldberg. Are Women Prejudiced against Women? *Society*, 5(5):28–30, 1968.

Claudia Goldin and Cecilia Rouse. Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians. *American Economic Review*, 90:715–741, 2000.

Benjamin Golub and Matthew O. Jackson. How Homophily Affects the Speed of Learning and Best Response Dynamics. *FEEM Working Paper*, 2012.

Catherine Good, Joshua Aronson, and Michael Inzlicht. Improving Adolescents' Standardized Test Performance: An Intervention to Reduce the Effects of Stereotype Threat. *Journal of Applied Developmental Psychology*, 24(6):642–662, 2003.

Catherine Good, Joshua Aronson, and Jayne Ann Harder. Problems in the pipeline: Stereotype threat and women's achievement in high-level math courses. *Journal of Applied Developmental Psychology*, 29(1):17–28, 2008.

Stefanie Gosen. Social Desirability in Survey Research: Can the List Experiment Provide the Truth? *Ph.D dissertation, Philipps-Universität Marburg*, 2014.

Alexander R. Green, Dana R. Carney, Daniel J. Pallin, Long H. Ngo, Kristal L. Raymond, Lisa I. Iezzoni, and Mahzarin R. Banaji. Implicit Bias among Physicians and Its Prediction of Thrombolysis Decisions for Black and White Patients. *Journal of Internal Medicine*, 22(9): 1231–1238, 2007.

Seth Green, Donald P. Green, Kulani Dias, and Betsy Levy Paluck. The Contact Hypothesis Re-examined. *In progress*, forthcoming.

Anthony G. Greenwald and Mahzarin R. Banaji. Implicit Social Cognition: Attitudes, Self-esteem, and Stereotypes. *Psychological Review*, 102(1):4, 1995.

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

Anthony G. Greenwald, Mahzarin Banaji, and Brian A. Nosek. Understanding and Using the Implicit Association Test: I. An Improved Scoring Algorithm. *Journal of Personality and Social Psychology*, 85(2):197–216, 2003.

Anthony G. Greenwald, T. Andrew Poehlman, Eric Luis Uhlmann, and Mahzarin Banaji. Understanding and Using the Implicit Association Test: III. Meta-analysis of Predictive Validity. *Journal of Personality and Social Psychology*, 97(1):17–41, 2009.

Jonathan Guryan and Kerwin Kofi Charles. Taste-Based or Statistical Discrimination: The Economics of Discrimination Returns to Its Roots. *The Economic Journal*, 123(572):F417–F432, 2013.

Daniel S. Hamermesh and Jeff E. Biddle. Beauty and the Labour Market. *American Economic Review*, 84:1174–1194, 1994.

Barton H. Hamilton, Jack A. Nickerson, and Hideo Owan. Team Incentives and Worker Heterogeneity: An Empirical Analysis of the Impact of Teams on Productivity and Participation. *Journal of Political Economy*, 111(3):465–497, 2003.

Andrew Hanson and Zackary Hawley. Do Landlords Discriminate in the Rental Housing Market? Evidence from an Internet Field Experiment in US Cities. *Journal of Urban Economics*, 70 (2-3):99–114, 2011.

L.R. Hartley and R.G. Adams. Effect of Noise on the Stroop Test. *Journal of Experimental Psychology*, 102(1):62, 1974.

James Heckman. Randomization and Social Policy Evaluation. In Charles Manski and Irwin Garfinkel, editors, *Evaluating Welfare and Training Programs*, pages 201–230. Harvard University Press, Cambridge, MA, 1992.

James J. Heckman. Detecting Discrimination. *Journal of Economic Perspectives*, pages 101–116, 1998.

James J. Heckman and Peter Siegelman. The Urban Institute Audit Studies: Their Methods and Findings. In Michael Fix and Raymond Struyk, editors, *Clear and Convincing Evidence: Measurement of Discrimination in America*. Urban Institute Press, 1993.

Madeline E. Heilman and Richard F. Martell. Exposure to Successful Women: Antidote to Sex Discrimination in Applicant Screening Decisions? *Organizational Behavior and Human Decision Processes*, 37(3):376–390, 1986.

Günter J. Hitsch, Ali Hortaçsu, and Dan Ariely. Matching and Sorting in Online Dating. *The American Economic Review*, pages 130–163, 2010.

Jonas Hjort. Ethnic Divisions and Productions in Firms. *CESifo Working Paper Series*, 2013.

Karla Hoff and Priyanka Pandey. Discrimination, Social Identity, and Durable Inequalities. *The American Economic Review*, pages 206–2011, 2006.

Allyson L. Holbrook and Jon A. Krosnick. Social Desirability in Voter Turnout Reports: Test Using the Item Count Technique. *Public Opinion Quarterly*, 74:37–67, 2010.

Sander Hoogendoorn and Mirjam Van Praag. Ethnic Diversity and Team Performance: A Field Experiment. *Tinbergen Institute Discussion Paper 2012-068/3*, 2012.

Sander Hoogendoorn, Hessel Oosterbeek, and Mirjam Van Praag. The Impact of Gender Diversity on the Performance of Business Teams: Evidence from a Field Experiment. *Management Science*, 59(7):1514–1528, 2013.

Melissa A. Houlette, Samuel L. Gaertner, Kelly M. Johnson, Brenda S. Banker, Blake M. Riek, and John F. Dovidio. Developing a More Inclusive Social Identity: An Elementary School Intervention. *Journal of Social Issues*, 60(1), 2004.

Jim Hubbuck and Simon Carter. *Half a Chance?: A Report on Job Discrimination against Young Blacks in Nottingham*. Commission for Racial Equality, 1980.

Kurt Hugenberg and Galen V. Bodenhausen. Ambiguity in Social Categorization the Role of Prejudice and Facial Affect in Race Categorization. *Psychological Science*, 15(5):342–345, 2004.

Lakshmi Iyer, Anandi Mani, Prachi Mishra, and Petia Topalova. The Power of Political Voice: Women's Political Representation and Crime in India. *American Economic Journal: Applied Economics*, 4(4):165–193, 2012.

Nicolas Jacquemet and Constantine Yannelis. Indiscriminate Discrimination: A Correspondence Test for Ethnic Homophily in the Chicago Labor Market. *Labour Economics*, 19(6):824–832, 2012.

Franklin James and Steve W. DelCastillo. Measuring Job Discrimination by Private Employers against Young Black and Hispanic Seeking Entry Level Work in Denver Metropolitan Area. *Unpublished report. Denver: University of Colorado*, 1991.

David W. Johnson and Roger T. Johnson. *Cooperation and Competition: Theory and Research.* Interaction Book Company, Edina, MN, 1989.

James D. Johnson, Erik Whitestone, Lee Anderson Jackson, and Leslie Gatto. Justice is still not Colorblind: Differential Racial Effects of Exposure to Inadmissible Evidence. *Personality and Social Psychology Bulletin*, 21(9):893–898, 1995.

Marvin A. Jolson. Employment Barriers in Marketing. *Journal of Marketing*, 1974.

Benjamin F. Jones and Benjamin A. Olken. Do Leaders Matter? National Leadership and Growth Since World War II. *The Quarterly Journal of Economics*, 120(3):835–864, 2005.

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

Lee Jussim and Kent D. Harber. Teacher Expectations and Self-Fulfilling Prophecies: Knowns and Unknowns, Resolved and Unresolved Controversies. *Personality and Social Psychology Review*, 9(2):131–155, 2005.

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

James G. Kane, Stephen C. Craig, and Kenneth D. Wald. Religion and Presidential Politics in Florida: A List Experiment. *Social Science Quarterly*, 85(2), 2004.

J. Kang et al. Implicit Bias in the Courtroom. *UCLA Law Review*, 59(5), 2012.

Dean S. Karlan and Jonathan Zinman. List Randomization for Sensitive Behavior: An Application for Measuring Use of Loan Proceeds. *Journal of Development Economics*, 98(1):71–75, May 2012.

Phyllis A Katz. Intergroup Relations Among Youth: Summary of a Research Workshop. Research summary, Carnegie Corp, New York, 2000.

Phyllis A. Katz and Sue R. Zalk. Modification of Children's Racial Attitudes. *Developmental Psychology*, 14(5):447, 1978.

Kerry Kawakami, John F. Dovidio, Jasper Moll, Sander Hermsen, and Abby Russin. Just Say No (to Stereotyping): Effects of Training in the Negation of Stereotypic Associations on Stereotype Activation. *Journal of Personality and Social Psychology*, 78(5):871, 2000.

Kerry Kawakami, John F. Dovidio, and Simone Van Kamp. The Impact of Counterstereotypic Training and Related Correction Processes on the Application of Stereotypes. *Group Processes & Intergroup Relations*, 10(2):139–156, 2007a.

Kerry Kawakami, Curtis E. Phills, Jennifer R. Steele, and John F. Dovidio. (Close) Distance Makes the Heart Grow Fonder: Improving Implicit Racial Attitudes and Interracial Interactions through Approach Behaviors. *Journal of Personality and Social Psychology*, 92(6):957, 2007b.

Kerry Kawakami, Jennifer R. Steele, Claudia Cifa, Curtis E. Phills, and John F. Dovidio. Approaching Math Increases Math=Me and Math=Pleasant. *Journal of Experimental Social Psychology*, 44(3):818–825, 2008.

Giora Keinan. Decision Making under Stress: Scanning of Alternatives under Controllable and Uncontrollable Threats. *Journal of Personality and Social Psychology*, 52(3):639, 1987.

Valdimer Key. Southern Politics in State and Nation. *University of Tennessee Press*, 1949.

Florence Kondylis, Mushfiq Mobarak, Ariel Ben Yishay, and Maria Jones. Are Gender Differences in Performance Innate or Social Mediated? *Working Paper*, 2015.

Annabelle Krause, Ulf Rinne, and Klaus F. Zimmermann. Anonymous Job Applications in Europe. *IZA Journal of European Labor Studies*, 1(1):1–20, 2012a.

Annabelle Krause, Ulf Rinne, and Klaus F. Zimmermann. Anonymous Job Applications of Fresh Ph. D. Economists. *Economic Letters*, 117(2), 2012b.

Kory Kroft, Fabian Lange, and Matthew J. Notowidigdo. Duration Dependence and Labor Market Conditions: Evidence from a Field Experiment. *The Quarterly Journal of Economics*, 128(3):1123–1167, 2013.

James H. Kuklinski, Michael D. Cobb, and Martin Gilens. Racial Attitudes and the "New South". *The Journal of Politics*, 59(2):323–349, 1997a.

James H. Kuklinski, Paul M. Sniderman, Kathleen Knight, Thomas Piazza, Philip E. Tetlock, Gordon R. Lawrence, and Barbara Mellers. Racial Prejudice and Attitudes toward Affirmative Action. *American Journal of Political Science*, 41(2):402–419, 1997b.

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

Calvin K. Lai, Maddalena Marini, Steven A. Lehr, Carlo Cerruti, Jiyun-Elizabeth L. Shin, Jennifer A. Joy-Gaba, Arnold K. Ho, Bethany A. Teachman, Sean P. Wojcik, Spassena P. Koleva, et al. Reducing Implicit Racial Preferences: I. A Comparative Investigation of 17 Interventions. *Journal of Experimental Psychology: General*, 2014.

Kristin A. Lane, Mahzarin R. Banaji, Brian A. Nosek, and Anthony G. Greenwald. Implicit Measures of Attitudes. In Bernd Wittenbrink and Norbert Schwarz, editors, *Understanding and Using the Implicit Association Test: IV*, pages 59–102. Guilford, New York, 2007.

Kevin Lang. A Language Theory of Discrimination. *Quarterly Journal of Economics*, 101: 363–382, 1986.

Victor Lavy and Edith Sand. On The Origins of Gender Human Capital Gaps: Short and Long Term Consequences of Teachers' Stereotypical Biases. *NBER Working Paper No. 20909*, 2015.

Edward Lazear. Language and Culture. *Journal of Political Economy*, 107(6):S95–S126, 1999.

Robyn A. LeBoeuf, Eldar Shafir, and Julia Belyavsky Bayuk. The Conflicting Choices of Alternating Selves. *Organizational Behavior and Human Decision Processes*, 111(1):48–61, 2010.

Sophie Lebrecht, Lara J. Pierce, Michael J. Tarr, and James W. Tanaka. Perceptual Other-Race Training Reduces Implicit Racial Bias. *PloS one*, 4(1):e4215, 2009.

Michael R. Leippe and Donna Eisenstadt. Generalization of Dissonance Reduction: Decreasing Prejudice through Induced Compliance. *Journal of Personality and Social Psychology*, 67(3): 395, 1994.

Sheri R. Levy, Steven J. Stroessner, and Carol S. Dweck. Stereotype Formation and Endorsement: The Role of Implicit Theories. *Journal of Personality and Social Psychology*, 74(6):1421–1436, 1998.

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

John A. List. Informed Consent in Social Science. *Science*, 322(5902):672, 2009.

Penelope Lockwood and Ziva Kunda. Superstars and Me: Predicting the Impact of Role Models on the Self. *Journal of Personality and Social Psychology*, 73(1):91, 1997.

I. Lustig. The Influence of Studying Foreign Conflicts on Students' Perceptions of the Israeli-Palestinian Conflict. *Unpublished Masters Thesis, University of Haifa*, 2003.

Debbie S. Ma, Joshua Correll, Bernd Wittenbrink, Yoav Bar-Anan, N. Srirarm, and Brian A. Nosek. When Fatigue Turns Deadly: The Association between Fatigue and Racial Bias in the Decision to Shoot. *Basic and Applied Social Psychology*, 35:515–524, 2013.

Michael D. Martinez and Stephen C. Craig. Race and 2008 President Politics in Florida: A List Experiment. *The Forum*, 8(2), 2010.

Benjamin Marx, Vincent Pons, and Tavneet Suri. Homogeneous Teams and Productivity. *Unpublish manuscript, available at: http://www.novasbe.unl.pt/images/novasbe/files/INOVA_Seminars/Vincent_Pons.pdf*, 2015.

David A. Matsa and Amalia R. Miller. A Female Style in Corporate Leadership? Evidence from Quotas. *American Economic Journal: Applied Economics*, 5(3):136–169, 2013.

Margaret Maurer-Fazio. Ethnic Discrimination in China's Internet Job Board Labor Market. *IZA Journal of Migration 2012*, 1(12):1–24, 2012.

Allen R. McConnell and Jill M. Leibold. Relations among the Implicit Association Test, Discriminatory Behavior, and Explicit Measures of Racial Attitudes. *Journal of Experimental Social Psychology*, 37(435-442), 2001.

Frances McGinnity, Jacqueline Nelson, Pete Lunn, and Emma Quinn. Discrimination in Recruitment. *Equality Research Series*, 2009.

Shelby McIntyre, Dennis J. Moberg, and Barry Z. Posner. Preferential Treatment in Preselection Decisions according to Race and Sex. *Academy of Management Journal*, 23(4):738–49, 1980.

Katherine L. Milkman, Modupe Akinola, and Dolly Chugh. Temporal Distance and Discrimination: An Audit Study in Academia. *Psychological Science*, 23(7):710–717, 2012.

Conrad Miller. The Persistent Effect of Temporary Affirmative Action. *Job Market Paper*, 2014.

Markus M. Mobius and Tanya S. Rosenblat. Why Beauty Matters. *The American Economic Review*, pages 222–235, 2006.

Don A. Moore, Samuel A. Swift, Zachariah S. Sharek, and Francesca Gino. Correspondence Bias in Performance Evaluation: Why Grade Inflation Works. *Personality and Social Psychology Bulletin*, 36(6):843–852, 2010.

David Neumark. Detecting Discrimination in Audit and Correspondence Studies. *Journal of Human Resources*, 47(4):1128–1157, 2012.

David Neumark, Roy J. Bank, and Kyle D. Van Nort. Sex Discrimination in Restaurant Hiring: An Audit Study. *The Quarterly Journal of Economics*, 111(3):915–941, 1996.

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

Brian A. Nosek, Mahzarin Banaji, and Anthony G. Greenwald. Harvesting Implicit Group Attitudes and Beliefs from a Demonstration Web Site. *Group Dynamics: Theory, Research, and Practice*, 6(1):101–115, 2002.

John M. Nunley, Adam Pugh, Nicholas Romero, and R. Alan Seals. An Examination of Racial Discrimination in the Labor Market for Recent College Graduates: Estimates from the Field. *Working Paper*, 2014.

David Ong and Jue Wang. Income Attraction: An Online Dating Field Experiment. *Journal of Economic Behavior & Organization*, 111(C):13–22, 2015.

Philip Oreopoulos. Why Do Skilled Immigrants Struggle in the Labor Market? A Field Experiment with Thirteen Thousand Resumes. *American Economic Journal: Economic Policy*, 3 (4):148–171, 2011.

Frederick Oswald, Gregory Mitchell, Hart Blanton, James Jaccard, and Philip E. Tetlock. Predicting Ethnic and Racial Discrimination: A Meta-Analysis of IAT Criterion Studies. *Journal of Personality and Social Psychology*, 105(2):171–192, 2013.

Sasson Oz and Dov Eden. Restraining the Golem: Boosting Performance by Changing the Interpretation of Low Scores. *Journal of Applied Psychology*, 85:314–322, 1994.

Gerard Padro i Miguel. The Control of Politicians in Divided Societies: The Politics of Fear. *Review of Economic Studies*, 74(4):1259–1274, 2007.

Devah Pager. The Mark of a Criminal Record. *American Journal of Sociology*, 108(5):937–975, 2003.

Elizabeth Levy Paluck and Donald P. Green. Prejudice Reduction: What Works? A Review and Assessment of Research and Practice. *Annual Review of Psychology*, 60(339-367), 2009.

Rohini Pande. Can Mandated Political Representation Increase Policy Influence for Disadvantaged Minorities? Theory and Evidence from India. *The American Economic Review*, 93(4): 1132–1151, 2003.

Eleonora Patacchini, Giuseppe Ragusa, and Yves Zenou. Unexplored Dimensions of Discrimination in Europe: Religion, Homosexuality and Physical Appearance. *Unpublished manuscript: http://www.frdb.org/upload/file/FRDB_Rapporto_PATACCHINI.pdf*, 2012.

B. Keith Payne. Weapon Bias: Split-Second Decisions and Unintended Stereotyping. *Current Directions in Psychological Science*, 15:287–291, 2006.

B. Keith Payne, Yujiro Shimizu, and Larry L Jacoby. Mental Control and Visual Illusions: Toward Explaining Race-Biased Weapon Misidentifications. *Journal of Experimental Social Psychology*, 41(1):36–47, 2005.

Pascale Petit. The Effects of Age and Family Constraints on Gender Hiring Discrimination: A Field Experiment in the French Financial Sector. *Labour Economics*, 14(3):371–391, 2007.

Thomas F. Pettigrew and Linda R. Tropp. Does Intergroup Contact Reduce Prejudice? Recent Meta-Analytic Findings. *Reducing Prejudice and Discrimination*, 93(114), 2000.

Gerard A. Pfann, Jeff E. Biddle, Daniel S. Hamermesh, and Ciska M. Bosman. Business Success and Businesses' Beauty Capital. *Economic Letters*, 67(2):201–207, 2000.

Edmund S. Phelps. The Statistical Theory of Racism and Sexism. *American Economic Review*, pages 659–661, 1972.

E. Ashby Plant and Patricia G. Devine. The Active Control of Prejudice: Unpacking the Intentions Guiding Control Effects. *Journal of Personality and Social Psychology*, 96:640–652, 2009.

Devin G. Pope and Justin R. Sydnor. What's in a Picture? Evidence of Discrimination from Prosper.com. *Journal of Human Resources*, 46(1):53–92, 2011.

Canice Prendergast and Robert Topel. Favoritism in Organizations. *Journal of Political Economy*, 104:446–461, 1996.

Gautam Rao. Familiarity Does Not Breed Contempt: Diversity, Discrimination and Generosity in Delhi Schools. *Job Market Paper*, 2013.

M. Marit Rehavi. Sex and Politics: Do Female Legislators Affect State Spending? *Unpublished manuscript, University of Michigan*, 2007.

Stephen Reicher and Mark Levine. Deindividuation, Power Relations between Groups and the Expression of Social Identity: The Effects of Visibility to the Out-Group. *British Journal of Social Psychology*, 33(2):145–163, 1994.

Dennis Reynolds. Restraining Golem and Harnessing Pygmalion in the Classroom: A Laboratory Study of Managerial Expectations and Task Design. *Academy of Management Learning and Education*, 6(4):475–483, 2007.

Peter A. Riach and Judith Rich. Testing for Racial Discrimination in the Labour Market. *Cambridge Journal of Economics*, pages 239–256, 1991.

Peter A. Riach and Judith Rich. Field Experiments of Discrimination in the Market Place. *The Economic Journal*, 112(483):F480–F518, 2002.

Peter A. Riach and Judith Rich. An Experimental Investigation of Age Discrimination in the English Labor Market. *Annals of Economics and Statistics/Annales d'Économie et de Statistique*, pages 169–185, 2010.

Milton Rokeach. Long-Range Experimental Modification of Values, Attitudes, and Behavior. *American Psychologist*, 26(5):453, 1971.

Milton Rokeach. *The Nature of Human Values*, volume 438. Free Press, New York, 1973.

Dan-Olof Rooth. Obesity, Attractiveness, and Differential Treatment in Hiring: A Field Experiment. *Journal of Human Resources*, 44(3):710–735, 2009.

Dan-Olof Rooth. Automatic Associations and Discrimination in Hiring: Real World Evidence. *Labour Economics*, 17(3):523–534, 2010.

Robert Rosenthal. On the Social Psychology of the Psychological Experiment: The Experimenter's Hypothesis as Unintended Determinant of Experimental Results. *American Scientist*, pages 268–283, 1963.

Robert Rosenthal. Interpersonal Expectancy Effects: A 30-year Perspective. *Current Directions in Psychological Science*, pages 176–179, 1994.

Robert Rosenthal and Lenore Jacobson. Pygmalion in the Classroom. *The Urban Review*, 3(1): 16–20, 1968.

Robert Rosenthal and Donald B. Rubin. Interpersonal Expectancy Effects: The First 345 Studies. *Behavioral and Brain Sciences*, 1(3):377–386, 1978.

Laurie A. Rudman and Peter Glick. Prescriptive Gender Stereotypes and Backlash toward Agentic Women. *Journal of Social Issues*, 57(4):743–762, 2001.

Laurie A. Rudman and Matthew R. Lee. Implicit and Explicit Consequences of Exposure to Violent and Misogynous Rap Music. *Group Processes & Intergroup Relations*, 5(2):133–150, 2002.

Bruce Sacerdote. Peer Effects with Random Assignment: Results for Dartmouth Roommates. *NBER Working Paper No. 7469*, 2000.

Tamar Saguy, Nicole Tausch, John F. Dovidio, and Felicia Pratto. The Irony of Harmony Intergroup Contact can Produce False Expectations for Equality. *Psychological Science*, 20 (1):114–121, 2009.

Jeffrey W. Sherman, Angela Y. Lee, Gayle R. Bessenoff, and Leigh A. Frost. Stereotype Efficiency Reconsidered: Encoding Flexibility under Cognitive Load. *Journal of Personality and Social Psychology*, 75(3):589, 1998.

Margaret Shih, Elsie Wang, Amy Trahan Bucher, and Rebecca Stotzer. Perspective Taking: Reducing Prejudice towards General Outgroups and Specific Individuals. *Group Processes & Intergroup Relations*, 12(5):565–577, 2009.

Natalie J. Shook and Russell H. Fazio. Roommate Relationships: A Comparison of Interracial and Same-Race Living Situations. *Group Processes & Intergroup Relations*, 11(4):425–437, 2008.

Robert E. Slavin. How Student Learning Teams Can Integrate the Desegregated Classroom. *Integrated Education*, 15(6):56–8, 1977.

Robert E. Slavin. Effects of Biracial Leaning Teams on Cross-Racial Friendships. *Journal of Educational Psychology*, 71(381-387), 1979.

Robert E. Slavin. *Cooperative Learning: Theory, Research, and Practice*. Allyn & Bacon, 2nd edition, 1995.

Robert E. Slavin and Robert Cooper. Improving Intergroup Relations: Lessons Learned from Cooperative Learning Programs. *Journal of Social Issues*, 55(4):647–663, 1999.

Robert E. Slavin and Eileen Oickle. Effects of Cooperative Learning Teams on Student Achievement and Race Relations: Treatment by Race Interactions. *Sociology of Education*, 54(3): 174–180, 1981.

Richard E Snow. Pygmalion and Intelligence? *Current Directions in Psychological Science*, pages 169–171, 1995.

Keith E. Stanovich and Richard F. West. Advancing the Rationality Debate. *Behavioral and Brain Sciences*, 23(5):701–717, 2000.

Claude M. Steele and Joshua Aronson. Stereotype Threat and the Intellectual Test Performance of African Americans. *Journal of Personality and Social Psychology*, 69(5):797–811, 1995.

Walter G. Stephan and Krystina Finlay. The Role of Empathy in Improving Intergroup Relations. *Journal of Social issues*, 55(4):729–743, 1999.

Matthew J. Streb, Barbara Burrell, Brian Frederick, and Michael A. Genovese. Social Desirability Effects and Support for a Female American President. *Public Opinion Quarterly*, 72(1):76–89, 2008.

Henri Tajfel. Experiments in Intergroup Discrimination. *Scientific American*, 223(5):96–102, 1970.

Henri Tajfel and John C. Turner. An Integrative Theory of Intergroup Conflict. *The Social Psychology of Intergroup Relations*, 33(47):74, 1979.

Andrew R. Todd, Galen V. Bodenhausen, Jennifer A. Richeson, and Adam D. Galinsky. Perspective Taking Combats Automatic Expressions of Racial Bias. *Journal of Personality and Social Psychology*, 100(6):1027, 2011.

Takahiro Tsuchiya, Yoko Hirai, and Shigeru Ono. A Study of Properties of the Item Count Technique. *Public Opinion Quarterly*, 71(253-272), 2007.

Margery Austin Turner, Michael Fix, and Raymond J. Struyk. *Opportunities Denied, Opportunities Diminished: Racial Discrimination in Hiring*. The Urban Institute, 1991.

Eric Luis Uhlmann and Geoffrey L. Cohen. Constructed Criteria Redefining Merit to Justify Discrimination. *Psychological Science*, 16(6):474–480, 2005.

A.D. Van Knippenberg, A.P. Dijksterhuis, and Diane Vermeulen. Judgement and Memory of a Criminal Act: The Effects of Stereotypes and Cognitive Load. *European Journal of Social Psychology*, 29(2-3):191–201, 1999.

Bernd Wittenbrink, Charles M. Judd, and Bernadette Park. Evidence for Racial Prejudice at the Implicit Level and Its Relationship with Questionnaire Measurements. *Journal of Personality and Social Psychology*, 72:262–274, 1997.

Anita Williams Woolley, Christopher F. Chabris, Alex Pentland, Nada Hashmi, and Thomas W. Malone. Evidence for a Collective Intelligence Factor in the Performance of Human Groups. *Science*, 330(6004):686–688, 2010.

Bradley R.E. Wright, Michael Wallace, John Bailey, and Allen Hyde. Religious Affiliation and Hiring Discrimination in New England: A Field Experiment. *Research in Social Stratification and Mobility*, 34:111–126, 2013.

David S. Yeager, Dave Paunesku, Gregory M. Walton, and Carol S. Dweck. How Can We Instill Productive Mindsets at Scale? A Review of the Evidence and an Initial R&D agenda. In *White Paper for White House Meeting on "Excellence in Education: The Importance of Academic Mindsets"*, 2013.

David S. Yeager, Valerie Purdie-Vaughns, Julio Garcia, Nancy Apfel, Patti Brzustoski, Allison Master, William T. Hessert, Matthew E. Williams, and Geoffrey L. Cohen. Breaking the Cycle of Mistrust: Wise Interventions to Provide Critical Feedback across the Racial Divide. *Journal of Experimental Psychology: General*, 143:804–824, 2014.

John Yinger. Measuring Racial Discrimination with Fair Housing Audits: Caught in the Act. *American Economic Review*, pages 881–893, 1986.

John Yinger. Evidence on Discrimination in Consumer Markets. *Journal of Economic Perspectives*, pages 23–40, 1998.

Natalia Zinovyeva and Manuel Bagues. Does Gender Matter for Academic Promotion? Evidence from a Randomized Natural Experiment. *IZA Discussion Paper 5537*, 2011.

Asaf Zussman. Ethnic Discrimination: Lessons from the Israeli Online Market for Used Cars. *The Economic Journal*, 123(572):F433–F468, 2013.