Field Experiments on Discrimination

Prepared for the Handbook of Field Experiments

Marianne Bertrand

University of Chicago Booth School of Business, NBER, and J-PAL

Esther Duflo

MIT Department of Economics, NBER, and J-PAL

This version: March 26, 2015

1 Introduction

Black people are less likely to be employed, more likely to be arrested or shot while unarmed. 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 are more likely to graduate from college, and get better grades. While many in the media and public opinion would argue that discrimination is a key force in driving these patterns, convincingly establishing that it is indeed the case has proven much more difficult.

In its purest form, discrimination happens when a member of an identified group (women, Blacks, Muslims, immigrants, etc.) is treated differentially by virtue of belonging to this group, due to prejudice against, or distaste for particular group (**?**). When two people with different backgrounds are treated differently, such as when Blacks are more likely to be arrested, it is of course not necessarily because police officers are racist; the police officers might be observing other characteristics and behaviors that help explain the higher arrest rate of Blacks absent any kind of animus, and in fact any kind of consideration of race in the decision to arrest.

Economists distinguish “taste-based discrimination” from “statistical discrimination” **?**, where differential treatment is motivated by imperfect information: as a prospective employer, renter, or car salesman, tries to infer the characteristics of the person but only has access to very imperfect information, she uses all the information available, including group membership. So police officers may weigh race in their decision to arrest but they do not do so out of animus: they do so because they know race to be predictive of criminal behavior.

While taste-based discrimination is both unfair and inefficient (simply consider how it constrains the allocation of talent), statistical discrimination cannot be immediately viewed as inefficient and is more easily defendable under the utilitarian argument. Indeed economists would generally say that employers “should” statistically discriminate, as it is profit-maximizing, it is not motivated by animus, and it has been argued by some as “fair” since it treats people with the same expected (though not necessarily with the same actual) productivity identically.

In truth, the two notions may be more closely related than they appear. Among White Americans, the deep prejudice against Blacks that is uncovered by implicit association tests may be related to a vague notion that they are more likely to be in prison, and hence dangerous: is this taste or statistical discrimination? If a group never interacts with another one, or never experience them as employees or co-workers, they may be fairly ignorant about their quality. At best this would mean that employing, electing, or renting to them may seem more risky which, in the presence of risk aversion, is also source of statistical discrimination (**?**).

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 girl is not good at math, talented women may become discouraged and not become good at math. If teachers or employers assume that students of 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 blurring the lines between the different theories of discrimination.

A rich literature in economics, sociology, political science and psychology has used experiments (in the lab and in the field) to provide considerable evidence that discrimination, which we argue given the discussion above might adequately be broadly defined as differential treatment because of group membership, exists. The first part of this chapter is devoted to the various methods that have been used to measure such discrimination. We start by reviewing audit and correspondence studies. Correspondence studies represent by far the largest share of field experiments on discrimination so far; overall, they offer staggering evidence of pervasive discrimination against minority groups all around the world. We summarize this research and review some its key limitations. We also discuss a few alternative methods to measuring discrimination, many of them having been used in the psychology literature and developed for the lab; we argue that these alternative methods deserve more consideration by economists in interested in measures of discrimination for their field research. These alternative methods include Implicit Association Tests, Goldberg Paradigm experiments and List Randomization, and measures of willingness to pay to interact with minority group members.

If discrimination is indeed pervasive, what are its costs to the underpresented groups, and to society overall? The second part of this chapter reviews the work that addresses these issues. In particular, we explore the work that has studied the consequence of discrimination, from self-expectancy effects (e.g. about how the stereotypes and social identities that end up defining some groups and directly affecting their performance and behavior) to expectancy effects (e.g. how stereotypes and biases against minority groups may end up being self-fulfilling). We also review a broader literature on the costs (and benefits) of the limited diversity in organizations and groups that directly result from discrimination.

The third and final section of this chapter 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 psychological literature on both socio-cognitive and technological de-biasing strategies. We argue that a lot promising future field research 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.

2 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.[1]

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

The traditional answer to this key difficulty of measuring discrimination in observational data has been to saturate the regression with as many possible productivity-relevant individual-level characteristics as available, but ensuring that the researcher observes all that the decision-maker observes is a hopeless task.

Moreover, adding more and more controls to a regression could ultimately end up obscuring the interpretation of the evidence. Consider for example the labor market context. Minority workers might be best-responding to the discrimination they know to exist in the labor market 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 **?**, 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, a minority group member may decide to under-invest in education if they expect that they will not be able to obtain labor market returns for this education.

Audit or correspondence studies were developed to address these core limitations of the

[1] For a review of this earlier literature on the narrower topic of labor market discrimination, see chapter 48 by **?**.

regression approach to measuring discrimination. We review below both types of studies and discuss the extent to which they address these limitations of the regression approach, but also other new limitations they create.

2.1 Audit Studies

?, the best-known collection of audit studies exploring the extent of discrimination, describes the audit 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" (**?**). Note that this literature will typically be silent on whether the discrimination is statistical or taste based.

Results from the earliest audit studies can be found in **?**, **?**, **?**, **?** , **?**, Cross et al. (1990), **?**, **?**, and **?**.

Audit studies have been used in a variety of settings, not just the labor market. A well-known early example of the audit method is offered by **?**. 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. 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 made randomly. 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 that they could provide their own financing for the car at the beginning of the bargaining. In spite of the identical approach to bargaining, **?** finds that white males are quoted lower prices than white women or black (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 in the labor market is (**?**). That study 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. We designed the study so that a male and female pair applied for a job at each restaurant, and so that, on paper at least, 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 restaurants included in the study suggests that earnings are substantially higher in high-price restaurants, so that the apparent hiring discrimination has implications for sex 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 **?** who had matched pairs of individuals apply for entry-level positions, probing the impact of a criminal record. 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).[2] In total, 350 employers were audited. The effect of the criminal record was both statistically significant and meaningful in magnitude: 17% of actors with a supposed criminal record received a callback, compared to 34% of testers who said they had no criminal record. That is, an equally qualified candidate

[2] 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."

was rejected about half of the time if he had a criminal record. For black applicants, the effect was even larger: 5% of African-American ex-convicts received a callback, compared to 14% of Blacks with no record. Note that an African-American auditor without a criminal record was about as likely to receive a callback 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. An exception is **?** 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 **?** is that 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 the taste-based explanations for the data by combining the field data with results from a dictator game conducted in the lab with these card dealers. He finds that nonwhite 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 actual market transactions.

2.1.1 Limitations of Audit Studies

Many of these weaknesses of audit studies have been discussed in **?** and **?**. 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 **?**, 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 small belief by auditors that employers treat minorities differently can result in measured differences in treatment. This 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 interview process.

2.2 Correspondence Studies

Correspondence studies have been developed to address some these more obvious weaknesses of the audit method.[3] 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, pairs of resumes or letters of interest for a potential rental are sent in response to a job or rental advertisement. One of the resumes or letters of interest in each pair is assigned the perceived minority trait, and discrimination is estimated by comparing the outcomes for the fictitious applicants with and without the perceived minority trait. The most common, but not exclusive, 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 (see below) limited to measuring call-backs by employers or landlords in response to the mailed or emailed fictitious application.

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

[3] We discuss in section ??? other weaknesses that are shared by the correspondence studies. We also discuss in that section added weaknesses of the correspondence tests compared to the audit method.

discrimination.

Although **?** call correspondence tests a “significant methodological advance,” and a review of discrimination in the marketplace published about fifteen years ago (**?**) discussed only observational and audit studies, the method is actually not that new. Fictitious applications and resumes have been sent to employers in order to uncover racial or religious discrimination nearly half a century ago.[4] However, the number of correspondence studies in economics has greatly increased following **?**. In this paper, they study race discrimination in the labor market by sending fictitious resumes in response to help-wanted ads in Boston and Chicago newspapers. To manipulate perceived race, they randomly assign very White-sounding names (such as Emily Walsh or Greg Baker) to half the resumes and very African-American-sounding names (such as Lakisha Washington or Jamal Jones) to the other half. To study how credentials affect the racial gap in callback, they also experimentally vary the quality of the resumes used in response to a given ad. Higher-quality applicants have on average a little more labor market experience and fewer holes in their employment history; they are also more likely to have an e-mail address, have completed some certification degree, possess foreign language skills, or have been awarded some honors. In practice, they send four resumes in response to each ad: two higher-quality and two lower-quality ones. They randomly assign to one of the higher- and one of the lower-quality resumes an African-American sounding name. In total, they responded to over 1,300 employment ads in the sales, administrative support, clerical, and customer services job categories and send nearly 5,000 resumes. They found that white names receive 50 percent more callbacks for interviews, and that the call-back rate *increased* with resume quality, which they take as an indication that statistical discrimination is unlikely to be the whole story.

2.2.1 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 resume. Below, we review some of these studies in more details, focusing in particular on those that have attempted to go beyond simply documenting whether or not

[4] See **?**, **?**, **?**, **?**, and **?** for early studies. One caveat is that some of these studies fail to fully match skills between minority and nonminority resumes.

Table 1: Labor market correspondence studies

Paper	**Country**	**CVs / apps**	**Vacancies**	**Effect (Callback ratio)**	**Theory**
? Trait: Ethnicity; Attractiveness	Peru	4,820	1,205	White-to-indigenous ratio: 1.8 Low attractiveness hurts white females	No
? Trait: Unemployment duration	Sweden	8,466	-	Employed to long-term unemployed: 1.25	No
? Trait: Arabic name	Netherlands	636	-	Dutch-to-foreign: 1.62 (unconditional ratio) No difference, if views held fixed	Threshold reading / attention discrimination
? Trait: Race	U.S.	9,396	-	White-to-black: 1.18 (unconditional)	Rejects statistical discrimination
? Trait: Unemployment duration	U.S.	3360	600	Employed-to-unemployed callback ratio: 1.47	No
? Trait: Ethnicity (Roma, Asian, Turkish)	Czech Rep. and Germany	274 (Czech R.) 745 (Ger.)	-	Czech-to-Vietnamese: 1.34 Lower requests for CVs if cand. is Turkish	Yes: Attention discrimination
? Trait: Religion / ethnicity	U.S.	6,400	1,600	White-to-Muslim: 1.58	Limited (secularization theory and religious stratification theory)
? Trait: Unemployment duration	U.S. (largest 100 MSAs)	12054	3,040	1 log point change in unemployment duration: 4.7 percentage points lower call-back probability	Limited: Results in line with screening models
? Trait: Nationality (Turkish-sounding name)	Belgium	752	376	1.03 to 2.05 depending on the occupation	Tightness of the labor market
? Trait: Sexual orientation	U.S.	4,608	1,536	No effect	No
? Trait: Sexual orientation	Sweden	3,990	-	Heterosexual male to homosexual: 1.14 1.22 in favor of the heterosexual female	No
? Traits:Sexual orientation and religion	U.S.	4183	-	1.16 in favor of a Christian relative to a Muslim	No
? Traits: Sexual orientation and attractiveness	Italy	2,320	-	1.38 against homosexuals	No
? Trait: Immigrant (race/ethnicity)	Germany	1,056	528	German-to-Turkish: 1.29 (if no reference letter is included)	Consistent with stat. discrimination
? Trait:Ethincity	China	21,592	10,796	1.36 (Han-to-Mongolian) to 2.21 (Han-to-Tibetan)	Limited (rejecting tastes of consumers)
? Trait: Race / Nationality	U.S.	330	990	English-to-foreign names: 1.41 English-to-Black names: 1.46	Limited
? Trait: Age	Sweden	466	-	3.23 favoring younger candidates	No
? Trait: Nationality (and race)	Canada	12910	3225	1.39 to 2.71 (against Indian, Pakistani and Chinese applicants)	Not consistent with discrimination due to language fluency concerns
? Trait: Gender	Sweden	3,228	1,614	1.07 in favor of women	No
? Trait: Ethnicity	Australia	Above 4000	-	From 1.12 (White-to-Italian) to 1.68 (White-to-Chinese)	Limited (Customer concerns are partly confirmed)
? Gender	Australia	3,365	-	Male-to-female: 1.28 (female-dominated professions)	No
? Trait: Age	UK	1,000+	-	2.64 favoring younger candidates	No
? Trait: Attractiveness	Sweden	1,970	985	1.21 to 1.25 (but higher for some occupations)	No
? Trait: Nationality / race	Ireland	480	240	1.8, 2.07, 2.44 against Asians, Germans and Africans respectively	No
? Traits: Caste and religion	India	3,160	371	1.08 for software jobs (insig.) 1.6 for call-center jobs	
? Trait: Age	U.S.	App. 4,000	-	Young-to-older: 1.42	No
? Traits: Age, gender, number of children	France	942	157	1.13 to 2.43	No
? Trait: Ethnicity	Sweden	3,552	1,776	1.82 in favor of Swedish-sounding names	Yes (statistical vs. social distance)
? Trait: Race	U.S.	4,870	1300+	1.5 (1.22 for females in sales jobs)	Yes (various aspects of common theories are rejected)
? Trait: Race and religion	U.S.	300	-	White-to-black callback ratio: 4.2 for selling positions	Partial (evidence of customer-driven discr.)

differential treatment occurs based on perceived traits, and towards understanding which theory of discrimination may best fit the patterns in the data; one of our bottom line below though will be that, unfortunately, the studies have tended to be close replications of the original **?** for different populations or contexts. The literature has failed to push the correspondence methodology to design approaches to better test for various theories of discrimination.

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 American (Peru, where whites are compared to indigenous applicants **?**), Asia (China, where Han, Mongolian, Uighur, and Tibetan are compared **?**), Australia (where indigenous Chine and white are compared, **?**), Europe (where immigrants are compared to non immigrants in Belgium **?**, Ireland **?**, etc.). Immigrants and non immigrants are also compared in the US (**?**), where the call back rate for Albanian sounding names are is almost as low as that of the Blacks.

Various researchers have attempted to adapt the correspondence method to learn more about which theory of discrimination fits the patterns in the data best. The most common approach has been to try to provide corroborative evidence for (or against) statistical discrimination.

As discussed above, **?** sent four resumes in response to each job posting, two higher-quality ones and two lower-quality ones. They found that whites with higher-quality resumes receive nearly 30-percent more callbacks than Whites with lower-quality resumes. On the other hand, having a higher-quality resume had a smaller effect for African-Americans. In other words, the gap between Whites and African Americans widens with resume 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. **?** 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 resumes that they may use quick heuristics in reading these resumes. 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." In section XXX, where we discuss what affects discrimination, we return to this hypothesis, including potential tests and policy implications. These findings are replicated in **?**: Blacks received 14 percent fewer callbacks compared to whites and discrimination was not mitigated when productive characteristics were added to a résumé.

Some studies find results more supporting of statistical discrimination. **?** submitted 12,910 resumes, sent in response to 3,225 job postings in Canada. First, he compared (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 foreigners vs Canadian if they went to a Canadian university, and 1.43 for a foreign university. The call back rate fell dramatically if the job experience was mixed (1.85) or purely international (2.71). Moreover, candidate with Chinese last name used an English first name (Allen and Michelle Wang), their prospects on the job market improved. This raises the possibility that a large fraction of the "discrimination" is either statistical discrimination, or direct inference that the candidate's English is likely to be poor. However, even for foreign sounding names with Canadian higher education and labor market experience the call back rate is still 1.39, which is in line with the other studies comparing immigrants and non immigrants in the US. This is still a sizeable difference. The fact that employers make rationale inference on the employees on the basis of the resumes does not rule out the fact that there may remain implicit discrimination.

Perhaps even more striking, **?** sent out 528 pairs of applications in Germany to study the effect of a Turkish-sounding name. The German-to-Turkish callback rate was 1.29 when no reference letter is 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 the "soft information" present in the letter appears to remove the difference in call back rates even though other, harder information 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 if the degree (and nature of discrimination) depends on the nature of the profession. **?** sent paired applications for positions of IT professionals, drivers, construction workers, sales assistants, high school teachers, restaurant workers, accountants, cleaners, pre-school teachers, nurses. Overall, women are called back slightly more often then men. In male dominated professions, male have a slight (insignificant) advantage. **?** focused on female-dominated professions (waitstaff, data-entry, customer service, and sales jobs). The callback ratio in favor of women was 1.28.

A key question of interest would be the extent to which there is a bias against women with children, or against young women who may have children in the future. To our knowledge only one study, **?**, studies this aspect (in France). She sends resumes of women and men, with or without children, of age 25 or 37. She finds... It seems that more work would be warranted on this topic.

Caste and religion **?** use job characteristics for inferring the extent to which the upper caste discriminate against lower castes ("scheduled" and "other backward" castes). They sent resumes in response call center jobs software jobs in India. Upper-to-scheduled caste callback ratio for software jobs was just 0.8 (and insignificant), while the upper-to-scheduled caste callback ratio for call-center jobs was 1.37 (still insignificant). On the other hand, Upper-to-Other Backward Caste callback ratio for call-center jobs was 1.6 and significant.[5] This could be related to expected probability in call center (statistical discrimination), given the importance of a fluent English, which may not be fully conveyed in the resume. They find no discrimination against Muslims.

The potential impact of religion on job prospects in the US was explored by **?** Affiliation with a religion was signaled through student activities.[6] The Control-to-Atheist ratio was 1.15 (not significant), the Control-to-Catholic ratio was 1.15 (not significant) and the Control-to-Jewish ratio was 1.15 (not significant). However, the Control-to-Evangelical ratio was 1.27 (marginally significant) and the Control-to-Muslim ratio was 1.58, and significant at 1% level.

Unemployment spells More recently, researchers have applied the correspondence model to better understand patterns of labor market discrimination against the unemployed. The results appear to vary from study to study.

[5] Upper-to-Other Backward Caste callback ratio for software jobs was 1.08 and insignificant.

[6] A caveat, similar to the LBGT results below is that activism in a religious group signal more than just religion.

? 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 9 months, his or her callback probability fell by 20 percent.[7] **?** found 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 nonemployment up to six months, and once the candidate experienced 6 months of unemployment, interview requests fell by an extra 8 percentage points.

? relate these results to the inference problem of the managers. They replicate the result that longer employment duration reduces call-back rate, but also shows 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.[8]

Other charactetistics Resume studies are now using to try to detect discrimination in a number of less obvious domains.

A literature has tried to estimate discrimination against LGBT candidates. The problem in this case is to provide information that identifies a candidate at LGBT. In **?** (which was carried out in Sweden), gay identity was identified by the mention of a "spouse" of either gender in the cover letter, and voluntary work in (LGBT or not) organizations. Professions studies included those that are male-dominated (construction worker, motor vehicle driver, sales person, and mechanic worker), female-dominated (shop sales assistant, preschool teacher, cleaner, restaurant worker, and nurse), and neutral (a high school teacher). They find some mild evidence of discrimination (ratio of 1.14), which could be due to the nature of the signaling (e.g. working in gay pride, as opposed to the red cross, may be seen as a political gesture, not just revealing an identity).[9] In Italy, **?** finds higher discrimination (1.38) and in the US, **?** find none.

Age has also attracted some attention, and several papers (**?**, **?**, and **?**) find that younger

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

[8] This may also explain the finding in **?** since this particular study was carried out between March and November 2007.

[9] The same could of course be true of carrying a very black name. We will return to this issue below.

candidates are generally preferred. A fundamental issue is that it is hard to argue that age is not necessarily a proxy for productivity: the direct impact of age on productivity cannot be controlled by adding other variable **?** tries to control for physical fitness with hobbies (e.g. racquetball is supposed to indicate fitness) but this is moderately convincing.

Finally, physical appearance has also been studied: **?** studies obesity, **?**, investigate the Beauty premium.

2.3 Correspondence Studies in Other Settings

2.3.1 Rental Markets

Correspondence studies in the housing market have very much followed the same approach as correspondence studies 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 name is found in Sweden (**?**), (**?**), (**?**). Discrimination against Blacks and other minority ethnicity if found in **?**, **?** and **?**. Discrimination against immigrants (particularly muslims) is found in Italy in **?** and Spain in **?**. Discrimination against LBGT is found in **?**.

Another popular variation, parallel to the labor market literature is to provide more information (e.g. job, etc). Interestingly, unlike what we found in most of the labor market literature, positive information (e.g. "I do not smoke and I work full time as an architect") reduces 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) increases it.

2.3.2 Retail

The expansion of on-line platform allows researchers to look at the impact of race on retail. There are currently much such fewer studies, but the door is wide open for more such studies to be performed.

? studied 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 of paired emails

Table 2: Rental market papers

Study	Country	Inquiries	Effect	Theory
? Trait: Minority status (Arabic name)	Sweden	5,827	1.37 to 1.62 (against Arabic females and males respectively)	Limited (unemployed may have difficulty paying the rent)
? Trait: Race	US	14,237	White-to-Black: 1.19 (Alt.: Black-to-White: 0.84)	Yes
? 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]	Yes: Attention discrimination
? Trait: Race	UK	9,456	1.12 (sometimes smaller)	Not really
? Trait: Immigrant status; Language ability	Italy	3,676	1.24 to 1.48 (Italian to East European and Italian to Arab respectively)	Tested in effect of language ability by sending some ill-formed emails
? Trait: Minority status (Arabic name)	Sweden	1,032	1.44 (no information), 1.24 (detailed information about the applicant)	Limited
? Trait: Immigrant status	Spain	1,809	1.19 to 1.44	Limited: information provision helps
? Trait: Sexual orientation	Sweden	408	Straight-to-gay callback ratio: 1.27	No
? Trait: Immigrant (race/ethnicity/religion)	Sweden	1,500	Swedish-to-Arab male: 2.17	No
? Trait: Race / Ethnicity (Arab, African American)	U.S. (Los Angeles County)	1,115	1.35 (white-to-Arab), 1.59 (white-to-black), conditional on hearing back 1.98 unconditional	No

sent to sellers of second-hand cars. First, an inquiry coming from somebody with a Jewish-sounding name was 22% more likely to be receive a response than an email from an interested buyer with a minority-sounding name. That is, the researchers were able to show that minorities need to exert greater effort to engage in market transactions, where some amount of trust is required.

Second, a follow-up phone survey was used to elicit sellers' attitudes about minorities to tease out the potential mechanisms. 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 seem to explain the differential (average) treatment of Arabs. This pattern is consistent with statistical discrimination.

? reports evidence from peer-to-peer lending sites. They find that loan listings with blacks in the attached picture are 25 to 35 percent less likely to receive funding than those of whites with similar credit profiles.

2.3.3 Academia

? 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 1 week; students' names signaled their race (Caucasian, African American, Hispanic, Indian, or Chinese) and gender. When the requests were to meet in 1 week, Caucasian males were granted access to faculty members 26% 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 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).

2.4 Beyond the resumes

Employers have access more information than just the resumes. A very small number of studies enrich the treatment but allowing employers to search for more (and different) information that would typically be available in a resume.

Given the increasing popularity of online social networks, the contribution of **?** is particularly interesting. They employ the correspondence method by submitting realistic applications to job posting and they extend their experiments by creating either personal websites of social networking profiles, 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 sexuality. The question the paper is asking is whether extra information available on line but not on the resume leads to discrimination: Would applicants whose identity is not revealed in the application, but who appear to be Muslim (vs. Christian) or gay (vs. straight) on a popular social network suffer unequal treatment?

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

The experiment finds that only a small fraction of employers use social media to conduct additional inquiry about job candidates. [10] Given the limited search efforts by employers, the effects of group membership are generally small. The total effect of trait manipulation is not statistically significant: 12.6% of applicants who appeared to be Christian received callbacks, compared to 10.9% of candidates who appeared to be Muslim. About 10.6% candidates who appeared to be straight males received callbacks, and the share of callbacks for seemingly gay males was nearly identical.

The strength of this type of study is that researchers are able to study the impact of traits that are traditionally not revealed on a resume. While some traditional correspondence tests have tried to signal religious affiliation or sexuality through "extra curricular activities" described on CVs, this type of disclosure might in itself be a signal on a resume (while it is entirely normal on

[10] Measuring the exact number of visits to a social networking profile is not possible for several reasons, but the authors estimate that at most one third of the employers tried to access the profile of the candidates.

a network). As we noted, perhaps some candidates are punished for signaling not only religion or sexual identity, but also a commitment to it as an identity. This may be what the employer is reacting to, rather than the religion or sexual identity per se.

? focus on the impact of religion and LGBT status, the effect of other interesting and until now mostly unexplored characteristics could be studied using a similar method. We mention only a couple of possible avenues for future research. 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 callbacks, 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, on-line field experiments offer a rich landscape for studying "what employers want."

2.5 Limitations of Correspondence Studies

As we discussed before, correspondence studies have helped address some key weaknesses of the audit study methodology. However, correspondence studies share some remaining weaknesses with the audit studies, and also introduce new concerns.

Both correspondence studies and audit studies can only inform us about 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 chapter 5 in **?** who challenge the use of newspaper advertisements in audit studies, referring to previous findings that most jobs are found through direct contract 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

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 apply 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, here, the correspondence studies are inferior to the audit studies. Most of the time, interview invitations or rental offers ("callback rates") are the only outcomes captured by simple field experiment (however, **?** were able to track transactions 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 does or does not get an appartment. All of this can be achieved in an audit study. 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"? **?** 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 that are typically 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. Researchers that carry out audit studies and correspondence studies are using this scarce resource without the involved parties' consent. A positive take on this ethical issue is **?** 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 non-members of the scientific community would probably offer a difference perspective. (In fact, List 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.).[11]

Another under-appreciated 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 become then become less intense, which might be beneficial for real jobless candidates (but potentially have a negative effect on the employers). This is more likely to be an issue when there are few responses to a given add (e.g. rental market) and thus the experimental add can bias the ratios.

A more subtle criticism by **?** is recently revisited in **?** relates to a more subtle. **?** show that a troubling result emerges in audit or correspondence studies because the relevant treatment 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 an interview) if the standard is met. The intuition for the critique is then as follows. Consider the simplest case in which the only difference between blacks and whites is that the variance of unobserved productivity is higher for whites than for blacks, for example. The correspondence study makes the two groups equal on characteristic X_1. The correspondence study does not convey any information 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 observeX_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

[11] The method of correspondence studies has been taken to the dating market (e.g XXXX). We do not study these contribution 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 seem particularly acute. As a conceptual aside, it is also not at all clear that one needs to send fictitious profile on dating web sites, as it is already possible for the researchers to observe exactly the same observation that the decision maker has when making a decision. There is thus no "unobserved" variable biasing the analysis and no information to be gained from fictitious resumes. The exercise can be performed with observational data (See **?**; **?**; **?**). This makes the ethical concern particularly salient.

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 the standardization is at a high value of X_1; in the latter 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, **?** 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, **?** shows that if a correspondence study includes observable measures of variation in 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 unobservable. Neumark applies this to /citet* Bertrand:2004vu correspondence study, and finds in the context of their data the evidence for race discrimination that adversely affects blacks than is obtained is indeed biased upwards when differences in the variances of the unobservable are ignored. Neumark explains how his method can be easily implemented in any future correspondence study. All that is needed is for the resumes or applicants to include some variation in characteristics that affect the probability of being hired. This is different from what is often done in designing these studies, where researchers try to create a pool of equally-qualified applicants. In contrast the researcher must intentionally create resumes of different quality. Once she confirms that a set of productivity-related characteristics on the resumes affected hiring outcomes, it should then be possible "conditional on an identifying assumption that has testable implications" to detect discrimination.

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 that is necessary for identification in this case. Finally, to solve the identification problem highlighted by **?** 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 do not vary with group membership (for example, race). This identifying assumption has testable implications in the form of overidentifying restrictions.

Finally, it is remarkable that after 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 by the recipient. For example, **?** suggest that black names may "provide a useful signal to employers about labor market productivity after controlling for information on the resume." This is clearly true for age, as we noted. But this may be true for black names if the choice of a black name is a political statement by the parent, accompanied by a different attitude towards schooling and obedience. More generally, as we already mentioned several time, even if employers do not in general 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, but not in the resume (**?**).

The only approach that has been repeatedly used by researchers to try to separate statistical discrimination 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 resume or applications. As more productivity relevant information is included on the resume, average differences in unobservable between the minority applicant are reduced, and statistical discrimination should also be reduced. But it is clear that this remains a very indirect way to try to isolate taste-based discrimination among employers or landlords.

2.6 Beyond Call Backs

A very recent paper that breaks the mold of the typical correspondence study and deserves particular attention is **?**. This paper is remarkable in its ability to push the correspondence study methodology forward, think beyond the pure call back data and avoid the problem inherent with the "over-signaling" of particular trait, and refine our theories of discrimination.

The paper links two important ideas: attention is a scarce resource, and lack of information about individual candidates drive discrimination in selection decisions (e.g. 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 a selection decision. They argue that when a small share of applicants is above the bar, negative stereotypes are predicted to lower attention, while the effect is opposite when most applicants are above the bar.

They test for such "attention discrimination" in two field experiments: one in the labor market and one in the rental market, both carried out in the Czech Republic, where they can monitor the decision maker's information acquisition about applicants. They created personal websites for fictitious applicants and submitted rental applications in the Czech Republic, and job applications in Germany and the Czech Republic). The advantage of using a 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. The authors also ran a "no information treatment" to compare whether including a reference to a personal website changes the relative callback rates of whites to minorities (Asian or Roma). Hence, the study was able to show whether a minority-sounding name 1) leads to differential callbacks, 2) causes less or more search.

When no reference to a personal website was included then among these 451 rental inquiries sent in the Czech Republic, applicants who appeared to be white received nearly twice as many invitation to view a vacant apartment than individuals who appeared to be Roma or Vietnamese. When a link to a personal website was included in a rental query (n=762), landlords were more likely to click on it and to seek additional information if the applicant appeared to be belong to a minority group. When a white applicant included a personal website, there appeared to be no meaningful change in the number of invitations to view an apartment. When a minority applicant included a link to a personal website, his invitation rate increased by 8 percentage points. Hence, landlords paid more attention to the minority applicants and the availability of the additional information through the website helped those minority applicants. The patterns were quite different in the labor market. Based on their name alone, white applicants received a 75% to 180% boost in callbacks. Moreover, when employers read an application of a candidate who appeared to be Czech (rather than Asian), the probability that they will read his online resume increased by 34%. That is, attention allocation was reversed.

The data can 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" then 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 viewed a priori less favorably.[12] These results supports a role for endogenous attention, which magnifies the role of prior beliefs in discrimination. The model implies persistence of discrimination in selection decisions, even if information about individuals is available and there are no differences in preferences, lower returns to employment qualifications for negatively stereotyped groups, and for policy, the important role of the timing of when a group attribute is revealed.

More effort to go use infrastructure like resume study to observe richer outcomes, more tightly linked to specific theory, would 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.

3 Other Approaches to Measuring Discrimination

3.1 Implicit Association Tests

The implicit Association test (IAT) is a computer-based test that was first introduced by **?**. Developed by social psychologists Greenwald, Nosek and Banaji and 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

[12] Note, however, that in the Czech Republic, online resumes of candidates who appeared to be Roma were inspected at roughly the same rate as the resumes of candidates who appear to be white. In spite of this result, most the data in **?** is consistent with an endogenous attention model.

go “together”. [13] The typical IAT consists of 7 “phases,” including practice phases to acquaint the subjects with the stimuli materials and rules. Consider for example an IAS 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 of these phases, 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) need to be places on the other side of the screen. 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 need to be placed on the other side. The extent to which an individual dislikes black faces (in this case) is then measured by the difference in response time (measured in millisecond) between the stereotypical phases and the non-stereotypical phases. [14]. Two broad kind 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, e.g. Nosek et al (2002)). The first kind is called attitude IAT, and the second stereotype or belief IATs. Other types of IATs 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, Muslim, women etc.), or phenomena more akin to statistical discrimination (women and math, women and career, women and politics, etc.). There are also a number of meta-analysis, review articles, and criticisms papers. It is not the scope of this paper to review all of this literature. One important take-away from this literature though

[13] See **?** for an excellent introduction to IATs.

[14] In practice, of course, a number of choices must be made about how to use the data, and this is reviewed in **?**

is that IATs to do seem to be capturing something about attitudes, perhaps more accurately than self-reports. **?** conducts a meta-analysis of 122 research reports using the IAT. They show that there is a strong correlation between implicit and more standard explicit measures. However, the IAT appear 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). For example, implicit bias predicts a more negative judgment of ambiguous actions by a black target (**?**), as well as more negative non-verbal behavior (less peaking time, less smiling, etc) during an interaction with a Black subject (**?**). 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 (**?**). Only a few studies have investigated whether these differences in implicit attitudes are associated with different behavior in the field. Doctors with stronger anti-black implicit attitudes were less likely to prescribe thrombolysis for myocardial infarction to African American patients, compared to white patients (**?**). **?** tried to relate the behavior of recruiters in a correspondence study in Sweden (focusing on Arab-Muslim vs Christian) to recruiter-level measures of implicit discrimination they collected later. They unfortunately were only able to interview 26% 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. IATs have been subject to a number of criticisms and questions, mainly regarding their interpretation. First, to the extent 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? Do IATs really identify prejudice? What does it mean for someone not to feel that there are prejudiced against blacks but have their IAT showing automatic white preferences? (**?**). On this last question, **?** would 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, it might be very wrong to conclude that "if prejudice is not explicitly spoken, it cannot reflect a prejudicial feeling" (**?**). 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 exerting a value judgment or stating, in a sense, a fact of life? There is in fact considerable variability in the measured implicit attitudes, and the correlation between those and explicit attitudes, between the different IATs in similar domains, as well as between IAT attitudes and behavior, does seem to indicate that there is some signal about the individuals. This does not mean that the IAT can be considered to be a reliable measure of the attitude of any particular individual (at best it is measured with considerable noise). However, it means that IAT may be good measurement tools for the propensity for a group to discriminate towards each other. In this context, it is a little surprising that IAT have rarely been used by psychologist as outcomes: although there is very little discussion on the subject, and very little data, it seems that many social psychologists consider those attitudes to be "hard wired" and not easily influenced by environmental factors. This is however entirely an empirical question, and 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 it is indeed correlated with future behavior (which the psychologist have found), finding out if it can at all be affected by the economic environment seems important. In recent years, economics have started using IATs as dependent variables. For example, /citetpowerfulwomen 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). But our overall impression is that the technique is under-used in the field. /citetlane2007usingIAT 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 population who are not literate (this is what /citetpowerfulwomen use) and although they are more difficult in populations who have had no experience with computers /citetpowerfulwomen eventually only included subjects below a given age), they can be a very useful tool.

3.2 Goldberg Paradigm Experiments

Goldberg Paradigm experiments are laboratory versions of audit or correspondence studies. They are named after a 1968 experiment by **?**. In the original experiment, students graded written essays, which were identical except for the male or female name of their author. 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 (see **?** for a review and meta-analysis of the literature on resistance to woman leadership). 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 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, **?** do not find that, on average, women leaders are evaluated significantly more negatively than men leaders. However, they are in some circumstances, e.g. when the leadership was carried out in a masculine style (in particular when the leader was projected to be authoritative). This confirms 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, then 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, /citetpowerfulwomen 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 or pro-female. 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. (We will discuss the results below). /citetpowerfulwomen find that both men and women, but men more than women, tend to discriminate against female leaders (additional results from this study are discussed below).

3.3 List Randomization

Like correspondence tests or Goldberg experiments, list randomization (also known as item count technique, or 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 of false (e.g. I had coffee at breakfast; I like popcorn).[15]Then, a randomly selected group of people is asked a potentially controversial statement (e.g. "I would be upset of 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 to be affected by social desirability bias. Early applications of list randomization to measure discrimination are **?**and **?**. Both studies found considerable racial prejudice in the 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 based on answers to the question of whether a respondent would be comfortable with "A black family moving next door." Likewise, **?** show that stated discrimination against LGBT 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, respondent were 67% 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. A few papers have used the method to elicit attitudes towards presidential candidates. **?** find no discrimination against a Jewish presidential candidate (Lieberman). **?** find that few whites in Florida seemed distressed by the possibility of having a Black President. However, list randomization revealed much more

[15] 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.

opposition towards the idea of a female president than opinion polls (**?**). 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, **?** found that 63% of the estimates for socially undesirable behaviors were significantly larger when elicited through list randomization. On the other hand, responses on nonsensitive behavior tend to be more similar (e.g. **?**).

The list randomization method is however not without issues. As we already alluded to earlier, there is a fundamental tension between precision (which would require having statements to which everybody says yes or no to) 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. **?** have also shown results tend to systematically depend on how many non-controversial statements are included in the list, although the opposite was found in **?**.

In summary, this is a promising method to measure discrimination as it less subject than social desirability bias, but since few economists have used it (see **?** for another application), 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) (**?**). 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.

3.4 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 tends to be based on binary measure (interview or not, hire or not). The closest to do that, until recently was a literature on the "beauty premium", motivated by **?**'s (**?**) finding that workers with better than average looks earn 10

% to 15% higher wages. Analogous to the black-wage race gap, the beauty premium could be do to the fact that more beautiful workers are more productive, say because consumers prefer to interact with beautiful people (**?**), (**?**), more confident, or simply wrongly believed by their employers to be more productive. **?** set up a laboratory experiment where "employers" must hire "workers" to perform a maze-solving task. After a practice test (which is recorded and becomes the digital "resume" of the worker) and a question where the workers need to assess how many mazes they can solve in 15 minutes, each worker is matched to 5 employers, who sees either (1) just the resume or (2) the resume and a photo or (3) the resume and a phone interview or (4) the resume plus an interview, plus the photograph. The employers in turn see five workers, and for each of them decides how many maze they think the worker can solve. This estimate contributes to their own payment. It also enters into the calculation of the actual wage of most of their workers. They 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 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 beauty premium by comparing treatments, they estimate that 15% is due to the confidence channel and 40% 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 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 Moebius and Rosenblatt's footstep is **?**. Rao 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 later, 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, he 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 that were 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 is very high stake). After mixing for 15 minutes, they watched a series of one-on-one sprints (most of them pitching a poor student against a rich one) and were then asked to indicate on a worksheet which of the two they want as teammate on 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 don't like them), students then had to spend two hours socializing with their teammate; this 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% of the cases where the poor students is the fastest, rich students prefer to pick the rich student as a teammate anyway. [16] Discrimination does decline as the stakes increase: discrimination falls from 35% with the lowest stake, to 27% with the intermediate stake, and 5% in the highest stake. Fitting a structural model so the data, Rao estimates that, for students without prior exposure to poor classmate, 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.

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

4 Consequences of Discrimination

4.1 Self-Expectancy Effects

4.1.1 Stereotype Threat and Underperformance

Models of statistical discrimination explain the differential treatment of minority 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 average difference in productivity into account when deciding whether or not hire any African American job candidate, 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 minority individuals may suffer negative performance outcomes (such as lower test scores or less engagement with academics) because of the burden of the "stereotype," "stereotype threat" (**?**). The key conjecture is that the threat of being viewed through the lens of a negative stereotype can create an anxiety that disrupts cognitive performance.

In a seminal study, (**?**) 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 is not being used to measure their abilities – can significantly improve African Americans' performance, dramatically reducing the racial gap in performance.

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). [17]

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. This approach has encouraged stereotyped individuals to reject the idea that intelligence is a fixed trait and instead to adopt the mindset that intelligence is a quality that can be increased with hard work and effort.

For example, in a lab experiment, **?** 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 9 weeks following the intervention were significantly higher for Blacks in the malleable intelligence condition. Whites showed a similar, though statistically marginal, effect. This study showed that encouraging students to see intelligence as malleable (i.e., embrace an incremental theory of intelligence) can raise enjoyment and performance in academic contexts.

It is important to note that while the randomized interventions take place in the lab, out-

[17] While most of these lab studies have been conducted by social psychologists, a few have been performed by economists. For example, **?** ran a lab experiment with athlete 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, **?** present evidence from a caste priming experiment in Uttar Pradesh, India. 321 high-caste and 321 low-caste junior high school male student volunteers in village India participated in an experiment in which either their caste was not publicly revealed or it was made salient. 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 (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.

comes are measured 1) on naturally occurring tasks outside the lab and 2) quite a long after the interventions took place; both of these features are important strengths of this experiment compared to the standard "stereotype threat" lab experiments.

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 measurement is part of normal operations, have provided a natural setting for this much field research.

/citet*good2003 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 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 minority and low-income adolescents – in the experimental conditions earned significantly higher reading standardized test scores than students in the control condition.

Also, /citet*good2008 conducted a field experiment where they explore stereotype threat and its negation in high-level college math courses that typically serve as gateway courses for 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 the GRE. 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, /citetcohen06 reduced the blackwhite GPA gap among lowincome middle school students by affirming the students' self-concepts (and presumably inoculating them from stereotype threat) at the beginning of the school term.

4.1.2 Identity and Preferences

The "stereotype threat" literature can be viewed as part of a broader literature that has been interested in how self-identity considerations may affect behavior and preferences of minority groups and ultimately may perpetuate gaps in economic outcomes between groups. 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).

In the backdrop of a large literature in social psychology that has tested the validity of the self-categorization theory and the cognitive mechanisms through which it operates (see for example **?**, **?**, and **?**) 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, **?** explore the effect of racial and gender category norms on time and risk preferences. They study in a laboratory setting 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 in temporarily making more salient ("priming") a certain social category (as is done in the "stereotype threat" 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 and where subjects are asked to identify their gender and whether they are 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 by **?**. The question under study here is whether gender identity priming affects subjects' level of altruism. The experiment consists in comparing behavior in a dictator game for subjects whose gender identity has been primed versus not. The results indicate that the priming does affect behavior (with women being more generous) 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 "stereotype threat" literature discussed above. It seems worthwhile for future research to consider such work. For example, interventions might be designed to play a "default" social identity that may be counter-productive 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 "father" identity 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).

4.2 Expectancy Effects and Self-Fulfilling Prophecies

4.2.1 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 (**?**; **?**). 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 his 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.[18] 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, **?** 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 much faster than the rest, as measured by IQ points. In fact, this set was randomly selected. The main outcome measure was an IQ test (Harvard Test of Inflected Acquisition), 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 learned more than control students in the same classes, 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 the "Pygmalion effect."

The Pygmalion effect in the classroom was subsequently studied intensively (see **?**; **?** for reviews), and by now considered a robust feature teacher-student relations.

[18] 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. **?** summarized a dozen experimenter-effect studies and wondered whether similar interpersonal expectation effects occur among physicians, psychotherapists, employers, and teachers.

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. **?** and **?** provide a review of much of the work in this literature. For example, **?** replicated the original design and results of **?** in the Israeli Defense Forces. But they also concluded based on additional survey work to complement the randomized control trial that leadership behavior was a key mediator in generating the Pygmalion effect (**?**).

Also using the Israeli Defense Forces as a field, **?** 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 7-week clerical course divided into 5 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. Another 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 directly. 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 relief instructors, reflecting the possible durability of expectancy effects. Finally, **?** 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. This has been dubbed the "Golem effect" by psychologists.

There have been much fewer studies on the Golem effects than the Pygmalion effects given the trickier ethical issues associated with lowering leaders' expectations (**?**). This challenge has led to research designs that are not quite as "clean" as those used to demonstrate the Pygmalion effects. For example, **?** 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 failed to introduce lower supervisory expectations or to compare the results with those of a control group, thus failing to follow the methodology applied in the majority of work designed to test the Pygmalion effect (e.g. **?**).

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, **?** studied expectation effects among physical education student-teachers. They found that pupils about whom they imparted high expectations to the instructors performed best (e.g. 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."

4.2.2 Endogenous Responses to Bias

While the Pygmalion and Golem effects demonstrate the self-fulfilling nature of leaders' expectations about performance on performance, there are not directly tied to discrimination. Are leaders' biases against minority groups also endogeneously affecting the performance of these groups? Two very 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 demonstrate to that in **?** to demonstrate the relevance of self-fulfilling prophecies as an explanation for persistent differences in performance between minority and majority 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.

? study cashiers in a French grocery store chain. A sizable share of these cashiers is from North African and Sub-Saharan African origin. They assess whether cashiers perform worse on the days where they are assigned to a manager who is more biased against their minority group. They measure managers' bias towards workers from some origin using an IAT test. Because cashiers in these stores work with different managers on different days, and because cashiers have virtually no control over their schedule, they can 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.

? also report interesting complementary survey evidence to better understand mechanisms. They do not find that minority workers report disliking more working with biased managers, or that biased managers dislike them, or 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. **?** 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 track 3 cohorts of students from primary school to high school in Tel-Aviv, Israel. For identification, they 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 measure 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. 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 that subject national test administered during middle and high school, while it has a significant negative effect on girls. In addition, primary school math teachers favoring boys over girls encourage boys and discourage girls from engagement in advanced math courses offered in high school.

4.3 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 take care of their interests. In a standard median voter world, the under-representation of women or other minority groups in politics would not matter as much 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 under-represented groups in society will get worst outcomes (**?**) (**?**). This would also occur if the absence of a leader means that the under-represented 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 minority 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.[19]. There are also been a set of studies that exploit the random selection of places that have to elect a minority leader in India's local governments. Comparing villages that were randomly selected to receive either a male or female head, **?** find that female leaders spend more on goods that women prefer, compared to those that men prefer. **?** replicate the results over a longer time period, and using a data set 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 any more and women have (generally) left power. [20] **?** show that greater female representation (in local governments) is related to

[19] See **?**, **?**, **?**, **?** paper on SC reservation for MP in India in AEJ

[20] **?**compare places before and after reservation and does 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.

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 crimes, but rather greatest willingness to report such crimes. Finally, **?** further finds that village leaders from schedule castes invest more in scheduled castes hamlets.

4.4 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?

4.4.1 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.

? proposes that the roots of discrimination are communication difficulties across different groups (what he calls "language communities"). A similar argument is made by **?**. 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 counter-productive. 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 them, lots of valuable information will be not transmitted across groups. Arguments long these lines have been made, more or less formally, in the human resources and management literature. More formally, **?** show that, when agents in a network prefer to associate with those having similar traits (homophily), it may take 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 mean that diversity (and hence homophily) may in theory hurt or improve productivity (**?**, **?**).

While there is a large non-experimental literature on the impact on diversity on public good provision (See **?** for a review of the literature on diversity) and a sizeable lab experimental literature (e.g. **?**; **?**; , the field experimental literature is more nascent. There are nevertheless a few interesting recent papers we review below.

? and **?** 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. The experiment was run 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 nonDutch ethnicity varied from 20% to 90%). Students then had a year to choose their venture, elect officers, conduct meetings, produce, sale, make money, and liquidate. This is not a lab experiment: this plays out over a year, and the incentives to do well in the program are very high: their ability to graduates, their grades, and potentially some money, are all on the line.

There is a clear benefit to gender diversity in their 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. They 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: they find that the marginal effect of increasing diversity on performance is zero or perhaps even negative when there are least 50% Dutch. However, once Dutch are less than 50%, further increases in the share of minority 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.

? analyzes a natural experiment where a flower firm in Kenya randomly assigns 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 distributes flowers to a team of two downstream workers. The upstream worker earns w per flower packed, and the downstream worker is paid $2w$ per flower packed. He finds that, conditional on productivity, upstream workers distribute fewer flower 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 team, they give more flowers to the worker from their same ethnic group. Interestingly, the output gap between homogenous and ethnically mixed teams doubles 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, **?** randomly assign enumerators to pairs, and each pair to a supervisor. The job of the enumerator is 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 **?** experiment hints at a different potential negative impact of discrimination: discrimination may create room for cheating and corruption. In their setting, the co-ethnic supervisor was willing to let the enumerators cheat.

4.4.2 Discrimination and Corruption

? provide a theoretical analysis of the influence of favoritism on optimal compensation and extent of authority for 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 own's group (rather than less corrupt ones from another group) (**?**) (**?**). More generally, voters' preferences for a group may diminish the role of issues in campaign, and by implication the quality of government (**?**).

? provide non-experimental evidence of this effect: they show that in Sweden, after the social democratic party imposed gender balanced 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 takes places in Uttar Pradesh (?) . The rise in caste-based politics in Uttar Pradesh, India's most populous state, has been accompanied by a staggering criminalization of politics: on the eve of the 2007 election, 206 of the sitting members of the legislative assemblies had a criminal case pending against them (?). Prior 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, discussions). One type of campaign encouraged citizens to vote on issue, not on caste. There other to not vote for a corrupt candidate. They found that the caste campaign let to a reduction of the (reported) vote 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 above 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 (?). 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 seats is now 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 (2012) collected data on what happens in previously reserved places, they found that female incumbents whose seat 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 one 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.

4.4.3 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 (**?**) and for countries (**?**). 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: **?** and **?** examine the impact of the Norway 2006 law which mandated a gender quota in board. 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 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.

5 What Affects Discrimination?

5.1 Leaders and Role Models

When a group is discriminated against, there are fewer leaders from this group. 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 themselves 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 never observed success in practice. And third 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 even trying (XXreference for this?). In this world, 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 and points out the gap.

5.1.1 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 are more similar to them. [21]. This tendency is part of the rationale for requiring a certain fraction of women on corporate boards, or, in academia, in evaluation and promotion committees, etc.

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 bend over backwards to avoid being perceived as biased. In several observational studies, women were not inclined to judge other women more favorably than men.[22] In group decisions, there may also be a response of other members of the committee, who may try to "undo" any perceived agenda they perceive (rightly or wrongly) the minority group member 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 coauthors. **?** 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 men. Additional evidence in the study suggest that these results might be driven at least part by the fact that female evaluators tend to over-estimate the quality of male candidates.

? and **?** present interesting evidence from randomized academic evaluation committees in Spain and Italy, respectively. In both countries, candidates for promotion appear in front of

[21] **?** provide evidence from exam for entry into the judiciary in Spain that support the later effect; **?** show that the former effect applies in the case of academic promotions

[22] See **?** for an audit study in Australia, where they find no interaction between the gender on the resume and the gender of the recruiter; see **?** for similar evidence in the context of NSF proposal reviews, and **?** for referee reports.

a centralized committee to be qualified. Files are assigned to randomly composed committees. In the Spanish case, they 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, each additional female member decreases by around two percentage point the success rate of female applicants relative to that of male applicants. Analyzing the voting records they find both that 1) the same female candidate is voted on more harshly by females than by males, and 2) male committee members grade female candidates more harshly when there are women on their committee, perhaps because 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 fascinating (if a little depressing for the impact of this type of affirmative action). It would be interesting to see if is also applies to different settings, such as management or political decisions. One paper that makes some progress in this direction (although the contrast it focuses on is not experimental) is an audit study by **?**. He sent 3552 applications to 1776 jobs in Sweden, including applications to qualified positions, such as senior/high school teachers, IT-professionals, economists, and engineers, and compares, among other things, the call back rates of Swedish sounding and not Swedish sounding names according to the name of the CEO. He finds 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". This is consistent with all the evidence presented above.

5.1.2 Role Models and the Attitude of the Majority

Even if there is no direct effect of having women or minority on leadership decisions, it could still affect discrimination for minority because those individuals will function as role models.

In a working paper version of **?**, the authors propose a model where taste and statistical discrimination reinforce each other: suppose that there are strong taste (or social norms) against having a female leader. Then citizens have never observed one in action. This makes female leaders more risky as a group: even if citizen admit that they are equally competent on average, they have much more precise priors, and to the extent they are risk averse, this will lead them to avoid women leaders. This is of course re-inforce if they 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 people to experiment with women or minority leaders of colleagues (political leaders, member of boards, colleagues in academic department, students at top colleges) will have persistent impacts of reducing discrimination, even if it does not affect the underlying taste for the community, simply by affecting perceive competence. The impact will be reinforced to the extent that the image of what constitutes a good leader also evolves in response to what people see. **?**'s (**?**) "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 "good women." As people get to see many (not just one or two token) women who are strong leaders (but good people) or who have milder (but effective) leadership style, they may change their attitude towards female leaders. A potential drawback is the possibility of backlash against minority leaders, if there is a perception that they got there because of special treatment (**?**).

Empirically, in an observational study, **?** finds evidence of this persistence in the context of affirmative action in the US. US government contractors are forced to practice affirmative action. Miller finds that, after an establishment is deregulated because they are not a contracted any more, the black employment share continues to grow. **?** study a randomized natural experiment in the context of local electoral politics in India, and are able to provide more evidence for the mechanism underlying this persistence. They study a context where local village councils are randomly selected, by rotation, to be forced to elect female leaders. They show that, after a cycle of reservation (and even more when the same place happened to be reserved for a woman twice in a row), more women run, and are elected, on un-reserved seats. There could be many reasons for this (including the fact that women may be more willing to run, or networks of women may have been created), but they provide evidence that this is probably at least in part due to a change in attitude. They collect evidence on attitudes in various ways: with a Goldberg type experiment (as described in an earlier section), and with two IAT, one for like or dislike for female leader (a more "hardwired" attitude that their model takes no stance on) and one for stereotype associating women with domestic activities and men with leadership activities. They find that the experience with 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, in a rare example of using an IAT as an outcome variable, they show that 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: **?** also finds that women continue to be more likely to be elected after a seat was reserved for a while (in urban Maharastra). Finally, citetbanerjee2013 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 implication, and lab 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 female resume after reading an editorial documenting women's success in this type of job (**?**), 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, teachers, etc.). It would be interesting to establish whether such impact on the majority's attitude can be directly documented in some of these other important contexts.

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

The other effect of role models and trailblazers (perhaps even more emphasized than the first two) is by changing the attitudes of the minority group on their own ability to succeed. Seeing successful women or black may lessen stereotype threat (as discussed above), or the belief that society is rigged against people from my group so there is no point trying anyway. In both cases this could increase effort and lead to better outcomes for the minority, even without direct changes in the majority attitude (even though this could of course trigger subsequent change in the majority's beliefs and attitudes as well).

As in the case of the impact of exposure of the majority, there is both a descriptive literature and a laboratory literature on this question. The observational literature compares either outcomes (e.g. teen pregnancy) to naturally occurring exposure (e.g. black teachers), or direct measure of stereotypes (measured for example by IAT) to exposure. For example, **?** show that women who have ben 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 STEM increases the likelihood that girls will present themselves as interested in STEM (**?**, **?**). Interestingly, the answer seems to be that unless the role models are very carefully tailored, such exposure tends to make things worst (e.g. lowering girls' interest in STEM fields). Reviewing either literature is outside the scope of this chapter, but the mixed results of the lab experiments provide interesting clues about what the effect may be in the field.

Here again, the field experiment work appears to be more limited. **?** study the same randomized natural experiment for women leadership position in India, and look at the impact on girls' educational attainment and career aspirations. There is evidence of impact on parent's hopes for daughters. In never-reserved villages, parents were 14 percentage points (45%) less 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. This translated into a positive impact on education (girls were more likely to stay in middle school), which cannot be directly attributed to any direct action of the leader (because middle school are not under their jurisdiction). This is therefore strongly suggestive of a causality running from role model to aspiration to actual change in behavior.

Overall, this literature seems to us surprisingly thin, compare to the larger literature on "horizontal exposure" (e.g. roommates or classmates). 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 directly translate into behavioral changes for female students (or trainees) even absent any effect on aspirations. 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 affect minority groups' aspirations.

5.2 Intergroup Contact: Roommates and Classmates

? is often credited with the development of the contact hypothesis, also known as Intergroup

Contact Theory. The premise of Allport's theory states that under appropriate conditions interpersonal contact is one of the most effective ways to reduce prejudice between majority and minority group members. If one has the opportunity to communicate with others, they are able to understand and appreciate different points of views involving their way of life. As a result of new appreciation and understanding, prejudice should diminish. Allport's proposal was that properly managed contact between the groups should reduce these problems and lead to better interactions. Much of the psychology literature on the contact hypothesis has focused on lab experiments that have helped refine Allport's original theory. Evidence from a large set of laboratory experiments suggests that interactions with members of other groups in situations designed to reward cooperation can improve relations among group. Furthermore researchers have found that studies that included criteria for positive contact such as equal status and personal interaction were particularly effective at reducing prejudice (**?**) . But interactions in situations of competition have also been shown to exacerbate conflict, suggesting that the context in which contact happens matter, and contact can be sometime backfire. Yet it is ultimately difficult to assess the real-world relevance of these laboratory studies, both because they are typically short-term, and because it is unclear whether real-world situations resemble either the conflictual or the cooperative environments constructed in laboratory experiments. A few field studies have been able to bring the contact hypothesis to the real world. College roommates have turned out to deliver an ideal field context. **?** was the first to exploit the random assignment of college roommates at Dartmouth College for a study of peer effects, focusing on test scores. **?** leveraged a similar random assignment process of college roommates at Harvard to reflect on the power of contact in reducing stereotypical thinking and increasing empathy for minority groups. They studied 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 while white students assigned roommates from any minority group are more likely to continue to interact socially with members of other ethnic groups after their first year. The results suggest that mixing with members of other groups tends to make individuals more empathetic to these groups. What remains unclear from **?** is whether contact to a minority roommate reduced stereotype or bias. Empathy might increase even if bias is unaffected. **?** leveraged the same random assignment of roommates 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 Implicit Association Test. The reduction in stereotype is accompanied by a more general tendency to cooperate, as measured in a Prisoner's dilemma game, but muter effects on trust, as measured in a Trust game.[23] Related findings are reported in **?**. 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 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 and 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. Thus, the results suggest that interracial roommate relationships, although generally less satisfying and involving than same-race roommate relationships, do produce benefits. In a recent paper, **?** extends the field testing of the contact hypothesis from the dorm to the classroom. Using a combination of experimental and administrative data, Rao studies whether exposure of rich students (from 14 private schools in New Delhi) to poorer students affect (i) tastes for socially interacting with or discriminating against the poor; (ii) generosity and prosocial behavior; and (iii) learning and classroom behavior. Starting in 2007, some elite private schools in Delhi were required to offer places to poor students. Rao exploits this policy change to study the effects of being exposed to poor students. Core to his identification strategy is a comparison of outcomes for treated and non-treated student cohorts within a school. Rao also exploits a secondary 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 the schools, the number of poor children with names similar

[23] 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.

to a given rich student provides plausibly exogenous variation in personal interactions with a poor student. This identification strategy is obviously more appealing as a test of the contact hypothesis as it focuses more centrally on changes in personal interactions between students, and rules out other confounds (such as changes in teacher behavior, or changes in the curriculum). **?** finds that economically diverse classrooms cause wealthy students to discriminate less against other poor children outside school. As discussed in a prior section, 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 forced to reveal how they value they trade-off more-athletic poor students and less-athletic rich students. When the stakes are high - Rs. 500 ($10), about a month's pocket money for the older students - only 6% of wealthy students discriminate by choosing a slower rich student over a faster poor student. As the stakes decrease, however, discrimination increases. In the lowest stakes condition (Rs. 50), almost a third of students discriminated against the poor. Most interestingly, exposure to poor students at school reduces discrimination by 12 percentage points. Rao also conducts a second field experiment to see how taste-based discrimination is affected by being exposed to more poor students. 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 they required to attend the play date by 19%. Having a poor study partner (e.g. contact alone) explains 70% of the increase in this "willingness to play."

As indicated above, **?** also explores 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 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, if 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 detected 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.

One should however note that the magnitudes of the reduction in stereotype uncovered in the above studies above are quite small, except in **?**. More generally, while most of existing field evidence is consistent with contact being beneficial to reducing stereotypes, the lab results (as discussed above) also isolate situations under which contact may backfire. It would be nice for future work to address in the field what are much finer predictions from the social psychology literature as to when contact will help, and when it will hurt.

5.3 Socio-Cognitive De-Biasing Strategies

In the absence of direct contact, is it possible to *teach* individuals to overcome their stereotypes?

A field experiment attempts to do just that in Rajasthan, India. **?** set up a large scale randomized experiment designed (among other things) to test whether citizen can learn from other's experience about the quality of female leaders. This is an environment where, have we already shown, there is a large bias against the ability for women to be decision makers. Using high quality street theater troupes, they set up theater show followed by a discussion of the performance of local leaders a few weeks ahead of the 2010 Panchayat election. The idea, following up on the work we discussed previously that direct experience with a female leader does change attitude towards female leaders (and willingness to vote for one), was to see whether the process can be accelerated by providing citizens objective information that in fact, women and men are about equally good at carrying a key task of the local government The experiment took place in 382 panchayat. In randomly selected Panchayats, 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 were disaggregated by gender (as it turns out, women do about as well as men in the sample districts). There are two main results. First of all, the theater 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 in villages where the "gender neutral" campaign was run. For example, the incumbent vote share declines by 6 percentage point (or 60%!) in villages where the general campaign was run. Moreover, the vote share for the incumbent become more sensitive to past performance in places were the gender neutral campaign was run. Second, however, these effects disappear in places where the campaign introduce the "gender" theme: in those villages, there are very little effect of the intervention on any outcomes (including on the probability that a female runs or wins or the vote share for women). It is as if, when citizen understood that the campaign was about convincing them to consider women, they lost interest. This underscores the challenge of fighting discrimination in an environment where discrimination is rife.[24]

It is possible that this experiment failed because it did not pay enough attention to the structure of the bias, and to ways to overcome it. Over the last twenty years, social psychologists have designed and tested in the laboratory setting a series of strategies to reduce bias and stereotypical thinking. These efforts have yielded a number of strategies that have been shown to reduce implicit bias in the lab and in the short-term. Such cognitive strategies include: 1) taking the perspective of stigmatized others; 2) imagining counter-stereotypic examples; 3) training in negating stereotypical associations; 4) individuation. We consider them in turn.

There are now many studies attesting to the merits of perspective taking as a strategy for reducing intergroup bias. Perspective-taking involves stepping into the shoes of a stereotyped person. 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 can be very useful in assessing the emotional impact on individuals who are constantly being stereotyped in negative ways. Some studies have linked perspective taking to decreased activation and application of negative group stereotypes (**?**; **?**); others have shown that adopting the perspective of a particular outgroup target leads to more positive evaluations of other individual members of the target's group (**?**) and of the target's group as a whole (Vorauer & Sasaki, 2009). For example, **?** conducted a series of lab experiments examining the impact of perspective taking on several critical (but largely untested) intergroup 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). They included two different perspective-taking conditions in this experiment. Some participants

[24] Note that the effect of reservation in this sample on the probability that a woman runs or wins after the reservation is cancel is till 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 these types of intervention.

tried to imagine the Black man's thoughts, feelings, and experiences (perspective-taking – 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 (perspectivetakingself condition). After watching the video, participants completed a variant of the Implicit Association Test (IAT) that assesses automatic evaluations of Black Americans relative to White Americans. Subjects in both of the perspective-taking conditions exhibited significantly weaker pro-White bias than the control subjects.

Under counter-stereotypic imagining, an individual is asked to think of examples – either famous or personally known to the person – that prove the stereotype to be inaccurate. For example, if a person judges an African American male as lazy or incompetent, (s)he imagines Colin Powell or Eric Holder. **?** 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 and disliked White individuals, disliked Black and admired White individuals, or nonracial exemplars. Immediately after exemplar exposure and 24 hours later, they completed an Implicit Association Test that assessed automatic racial attitudes and 2 explicit attitude measures. Exposure to admired Black and disliked White exemplars significantly weakened automatic pro-White attitudes for 24 hour beyond the treatment but did not affect explicit racial attitudes. Experiment 2 provided a replication using automatic age-related attitudes. Also, **?** 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.

In **?**, lab subjects received extensive training in negating specific stereotypical thinking towards skinheads and elderly people. 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. 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 person with a nonstereotypic trait. **?** show that such training in negating stereotypes were able to reduce this stereotype activation. These results were obtained even when participants were no longer instructed to “not stereotype,” under predominantly automatic processing conditions, 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.[25]

? 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) ostensibly 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 of applicants involved two separate stages: judging the applicants along 16 different dimensions (8 stereotypically masculine traits like “risk-taker” and 8 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% chose a woman for the job. Bearing in mind that the gendered names and résumés were randomly mixed and matched for different participants, this can be interpreted as pure preference for giving the leadership position to a man. In contrast, among participants who had undergone counter-stereotype training, 61% chose a woman. [26]

? found that participants can also change their implicit biases and unreflective social behaviors by practicing “approach” and “avoidance” behaviors (approach training): White and Asian participants repeatedly pulled a joystick toward themselves when they saw black faces

[25] Two follow-up studies outside of Kawakami’s lab have partially replicated but partially qualified the original findings. First, **?** observed that the original studies confounded two sorts of training – the repeated affirmation of counterstereotypes and the repeated negation of stereotypes. Gawronski and colleagues thus split participants into two groups, all of whom saw the same overall set of face-word pairings, but instructed some to simply affirm the counterstereotypical pairings and instructed others to simply negate the stereotypical pairings. After 200 trials, participants who repeatedly affirmed counterstereotypes showed significant reductions in implicit biases, while those who negated stereotypes showed exacerbated implicit biases.

[26] 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% 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.

and pushed it away when they saw white faces. In pulling the joystick in, for example, it is as if participants are bringing the perceived image closer, or approaching it. This training significantly reduced participants' implicit bias on the IAT. In some cases, the images of the faces were"masked" and shown so quickly participants didn't notice them, and instead believed that they were just moving the joystick when they saw the words "approach" or "avoid." Significant reductions in implicit bias on the IAT were found in all conditions, regardless whether the meaning of the training was fully explicit or subliminal.

The same debiasing training method has also been shown to be promising (in the lab) to deal with the stereotype threat. **?** reported the beneficial effects for female undergraduates of repeatedly "approaching" math-related images ("e.g., calculators, equations"). Those who initially reported that they did not like math and were not good at it tended, after the training, to identify with and prefer math on implicit measures, as well as to answer more questions on a math test. A series of follow-up studies by **?** replicated these findings using a different training procedure, and with a 24-30 hour delay between the debiasing procedure and the math test. They also found that gender-math counter-stereotype training seemed more effective than approach training. Women subtly 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. Taken together, counter-stereotype and approach training seem to be effective procedures for debiasing ourselves along a number of key dimensions.

Individuating is another 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 (**?**) (**?**). **?** 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 5 sessions. In the individuation condition, subjects learned to discriminate between African American faces; specifically, they received "expertise training" with other-race faces "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. Concomitantly, subjects in the individuation condition, but not the categorization condition, showed a reduction in their implicit racial bias. Critically, 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 subject showed in their ability to differentiate African American faces.

A particularly exciting study in the socio-cognitive de-biasing area is **?**, 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 the 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. Among those that demonstrated effectiveness in this study were the three following, listed from most effective to least effective: 1) Shifting Group Boundaries through Competition: Participants engaged in a dodgeball game in which all of their teammates were Black while the opposing team was an all-White collective that engaged in unfair play. Participants were instructed to think positive thoughts about Blackness and recall how their Black teammates helped them while their White opponents did not; 2) Vivid Counterstereotypic Scenario: Participants read a graphic story in which they are to place themselves in the role of the victim who is assaulted by a White man and rescued by a Black man. Aiming to affirm the association that White = bad and Black = good, in each test of this intervention, the scenario was longer and enhanced by more detailed and dramatic imagery. Across three studies, this vivid counterstereotypic scenario substantially reduced implicit preferences among participants. 3) Practicing an IAT with Counterstereotypic Exemplars: Previous research established that exposure to pro-Black exemplars (e.g., Michael Jordan, Martin Luther King, Jr.) and negative White exemplars (e.g., Timothy McVeigh, Jeffrey Dahmer) decreases the automatic White preferences effect. This effective contest intervention used these counterstereotypic primes and combined them with repeated practice of IAT trials in which participants were to pair Black faces with Good and White faces with Bad.

All of this work has taken place in the lab. One concern on may have with this lab work is that it can only documents fairly short-term effects (up to 24 hours), and hence might be of limited relevance to the field. It would be interesting to devise field version of these tests for different population, and see whether they can work in practice. 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 short-term findings could be of real-world relevance, if the intervention where to take place just before one of those decisions times. What this work does not allow us to assess, however, is how these short-term findings would differ if the same person (e.g. an HR manager) was repeatedly exposed to the de-biasing strategy (e.g. every time he or she sits down to start reviewing resumes, or grading exams).

Some other de-biasing work in psychology has taken seriously these concerns about one-shot, short-term interventions and has asked whether related strategies can be built to produce enduring reduction in bias. Work by Devine and a series of co-authors is of particular interest. Devine proposes a habit-breaking analysis of prejudice reduction (**?**), which argues that overcoming prejudice is a protracted process that requires considerable effort in the pursuit of a non-prejudiced goal. This model 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 (**?**; **?**) argue that the motivation to break the prejudice habit stems from two sources. First, people must be aware of their biases and, second, they must be concerned about the consequences of their biases before they will be motivated to exert effort to eliminate them. Furthermore, people need to know when biased responses are likely to occur and how to replace those biased responses with responses more consistent with their goals. **?** 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 multifaceted nature of the intervention has conceptual parallels to approaches in several other areas, such as health behavior change, cognitive behavior therapy, and the fundamentals of adult learning. The effects of this multi-faceted habit-breaking intervention were studied over a three-month period. The key elements of the intervention were as follows. First, to ensure situational 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 such biases. 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 (e.g. the 4 strategies listed above: 1) taking the perspective of stigmatized others; 2) imagining counter-stereotypic examples; 3) training in negating stereotypical associations; 4) individuation) as well as seeking opportunities to engage in positive interactions with members of the minority group – e.g. contact) , enabling participants to flexibly choose the strategies most applicable to different situations in their lives. Following the manipulation, intervention group participants had lower IAT scores than control group participants. Moreover, the effects of the intervention on implicit race bias at 4 and 8 weeks were not systematically different from each other, indicating that the reduction in implicit race bias persisted throughout the 8-week interval. These data provide the first evidence that a controlled, randomized intervention can produce enduring reductions in implicit bias. The intervention however created no changes in either the participants' reported racial attitudes or their internal/external motivations to respond without prejudice. It did, however, affect participants' concern about discrimination and their awareness of their personal bias. Most interestingly, concern 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. Though effective overall, the complexity of the intervention results in ambiguity regarding which components are responsible for its various effects. Education may play a specialized role in increasing awareness and concern, but both education and training may be necessary to produce changes in implicit bias.[27]

[27] A different type of "education" is reporting of biased decision-making in the press. **?** showed that the racial bias of NBA referees disappeared after the media coverage of an academic study that revealed wide-spread bias in refereeing. The authors argue that referees adjusted their behavior because of "awareness" rather than institutional changes. But awareness is probably too broad and too limited at the same time as an explanation. It is broad because we do not know if the awareness of club owners, referees, or the audience that drove referees to call fouls more fairly. It is limited because the *cost* of biased refereeing probably rose after the coverage and publication of the original study. The term 'awareness' suggests that referees learned about their subconscious biases, repented, and changed their behavior voluntarily. (From the paper: "the evidence presented in this study suggests that the most likely mechanism through which the change in bias occurred is that the media reporting increased the awareness among referees about their own implicit racial bias and that this awareness led to a reduction in such bias.") This internal channel is certainly plausible; but is it not equally possible that external pressure (a more watchful audience) was responsible for the improvement in referees' behavior?

We are not aware of any field studies that have taken those insights from the socio-cognitive debiasing literature to the field and tested them in randomized control trial method.

There has been some attempt, however, to test these interventions in a lab-setting on experienced subjects. Some training interventions have been tested with police officers. **?** show that experience with simulated building searches – in which officers interact with actors, some of whom "attack" using weapons with non-lethal ammunition – does predict reduced bias.

In the same policing domain, there are also of a few recent initiatives exist that are ripe for evaluation, or in the process of being evaluated .

The US Department of Justice is funding the development of a curriculum for police staff that reflects on the Fair and Impartial Policing perspective. The Fair and Impartial Policing 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, sexual orientation, religion, socio-economic status and so forth. Officers are taught skills, inspired by the lab-tested methods above to reduce and manage their own biases. There are five Fair and Impartial Policing curricula customized to various audiences: Academy Recruits and/or In-Service Patrol Officers, First-Line Supervisors, Mid-Managers, Command-level Personnel (or Command Personnel and Community Leaders), Law Enforcement Trainers. The curriculum design teams are comprised of experts in the area of biased policing, police executives, first-line supervisors, officers, and community stakeholders. Additionally, and importantly, social psychologists from around the nation who conduct the research on human biases are members of this team. All five training programs have been implemented with the target audiences (recruits/patrol officers, first line supervisors, mid-level managers, command staff and law enforcement trainers) in multiple and diverse training environments, but to our knowledge they have not been subject to a rigorous evaluation.

Also, Goff heads an effort to help police departments drive down racial discrimination. Goff has conducted research on possible training methods by working with police departments in Chicago, Denver, Las Vegas, and many other American cities. Goff is collecting before-and-after statistics on things like traffic stops, arrests, and use of force by the officers he has worked with; however the results of this research have been published yet. There has also been much discussion in the recent years about how these methods could be used to de-bias judges, jurors or HR managers. For example, **?** discuss possibilities about how to import debiasing methods 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 countertypical exemplars or associations for participants in the trial process. Since judges and jurors are differently situated, we can expect both different effects and implementation strategies. For example, judges would be exposed to such vicarious displays regularly as a feature of their workplace environment. By contrast, jurors would be exposed only during their typically brief visit to the court. Especially for jurors, then, the goal is not anything as ambitious as fundamentally changing the underlying structure of their mental associations. Instead, the hope would be that by reminding them of countertypical associations, we might momentarily activate different mental patterns while in the courthouse and reduce the impact of implicit biases on their decisionmaking."

? provides the only attempt we are aware of to formally test for such cognitive de-biasing of jurors in a mock trial setting. The goal of the study was to deliver a de-biasing strategy that could be practically brought to the courtroom. This mock trial study used a 2 (defendant race: black or white) x 2 (victim race: black or white) x 2 (instructions: specialized implicit bias or control) factorial design. The goal was to examine the impact of a specialized implicit bias jury instruction on expressions of racial prejudice in juror decision-making. Participants, drawn from a nationally representative sample, eligible to be jury in the US, were asked to be the jury for a case of assault and battery which mixed evidence (intended to elicit a guilty or not guilty verdict roughly in mixed proportions). Two versions of jury instructions were constructed and videotaped for use in the experiment. Both versions of the jury instructions were delivered by an older white man in judicial robes from behind a judge's bench. The judge presented standard pattern jury instructions for reasonable doubt, battery causing serious bodily injury and self-defense in both instruction conditions. However, the experimental manipulation focused on the use of a specialized implicit bias jury instruction versus a control instruction of comparable length. Based loosely on a real jury instruction authors developed a specialized implicit bias jury instruction which incorporates concepts consistent with several promising bias-reduction strategies identified in the lab literature and summarized above: universality of bias, importance of awareness and encouragement of perspective taking. Following the jury instructions and presentation of the evidence in the mock trial, participants then answered a series of questions concerning their verdict preference (guilty/not guilty), confidence in their verdict preference (0% to 100%), assessments of the strength of case for the prosecution and for the defense, and sentencing recommendations. Three weeks later, participants completed an IAT and answered standard questions designed to measure explicit racism. Unfortunately, the results of this study are a mixed bag. Unexpectedly, the control conditions of the present study failed to generate the traditional patterns of juror bias, in which white mock jurors judge black defendants more harshly than white defendants (**?**). [28]

Without replicating this pattern of bias to establish a baseline against which participants in the experimental conditions could be compared, the authors were unable to fully examine the effectiveness of the specialized implicit bias jury instruction in reducing bias in juror judgments. However, they found no evidence to suggest that the specialized instruction produces a harmful backfire effect, even among those likely to feel threatened by it. There is also some subtle evidence of changes in the way that the strength of the defense case is evaluated, which are consistent with an impact of the debiasing but could also be due to chance. In summary, the field seems wide open for field experiments. We are confident that a large number of private and public organizations must have adopted or are considering adopting similar educational and training programs in socio-economic debiasing, adapted to their particular needs (such as for HR managers). It would particularly worthwhile for researchers in the future to strike partnerships with these organizations and rigorously evaluate the impact of this programing on organizational practices, and ultimately organizational performance.

5.4 Technological De-Biasing

%Other

? 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 im-

[28] Note that, while the original study seemed robust and has been replicated several times, the failure to replicate it in more recent data seem to generalize as well (At least three other research studies conducted after the present experiment (between November 2013 and January 2014) also failed to replicate the original juror bias effect, with college students and with two samples of web users recruited through Amazon Mechanical Turk (P. Ellsworth, personal communication, February 16, 2014).). This raises the possibility that the discussion of biases in the media has in itself contributed to debiasing potential juries. The increased salience of race and race norms in routine media communications about the American justice system could have primed participants to spontaneously self-monitor and correct for possible bias. If the latter is the case, a simple reminder to consider race and race norms may be sufficient to prompt jurors to engage in corrective action against expressions of bias in judgment.

proving 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), 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 (e.g., **?**; **?**).

Similarly, certain emotional states (anger, disgust) can exacerbate implicit bias in judgments of stigmatized group members, even if the source of the negative emotion hfas nothing to do with the current situation or with the issue of social groups or stereotypes more broadly (e.g., **?**, **?**). Happiness may also produce more stereotypic judgments, though this can be consciously controlled if the person is motivated to do so (**?**).

Also, 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 circumstances can adversely affect judicial performance (e.g., **?**; **?**; **?**). Specifically, situations that involve time pressure (e.g., **?**), that force a decision maker to form complex judgments relatively quickly (e.g., **?**, or in which the decision maker is distracted and cannot fully attend to incoming information (e.g., **?**; **?**) 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, the "shooter bias" discussed earlier has been shown to be exacerbated when respondents are tired (**?**), rushed (**?**), or cannot see well (**?**). Some of these circumstances are unavoidable during actual policing. However, any staffing and scheduling steps that minimize officer fatigue also could curb some of these racial disparities.

?'s (**?**) field study of sequential parole decisions made by experienced judges provide an another interesting illustration. Their sample is 1,112 parole board hearings in Israeli prisons, over a ten month period. These rulings were made by eight Jewish-Israeli judges, with an average of 22 years of judging behind them. Their verdicts represented 40% of all parole requests in the country during the ten months. Every day, each judge considers between 14 and 35 cases, spending around 6 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% to nearly zero within each decision session and returns abruptly to 65% after a break. The researchers argue that all repetitive decision-making tasks drain our mental resources. We start suffering from "choice overload" and we start opting for the easiest choice. For example, shoppers who have already made several decisions are more likely to go for the default offer, whether they're buying a suit or a car. And when it comes to parole hearings, the default choice is to deny the prisoner's request. The more decisions a judge has made, the more drained they are, and the more likely they are to make the default choice. Taking a break replenishes them. 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.

? study how one could build on this knowledge of what triggers System 1 vs System 2 thinking to help technology de-bias the courtroom. A lot of the possible strategies they discuss for the courtroom setting could naturally be applied to other real-world setting where biases in decision-making have been documented. For example, **?** 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 and courtrooms may 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 such greater ambiguity.

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 (**?**).[29] In general, the use of linear 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, **?** convincingly demonstrate racial gaps in attention allocation by HR managers. Once they see a minority name on a resume, they pay less attention to that resume. These findings strengthen the merit for thinking about the 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 for the outcomes of early pre-selection (we don't know of systematic evaluation of such strategies). 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 theory strengthens the case for suppressing the signals of a group attribute during early pre-selection. In this context it is intriguing that among policy-makers and private firms, the introduction of name-blind resumes has been receiving support during recent years. We now turn to this literature.

5.4.1 Blind Hiring Procedures

The large number correspondence studies has renewed interest for the possibility of using "blind" hiring procedures, which can be viewed as an extreme form of technological de-biasing. In some

[29] The value of linear models in hiring, admissions, and selection decisions is highlighted by research that **?** conducted on the interpretation of grades by graduate admission officers.

recruiting circumstances, the full hiring process can take place anonymously. **?** famously showed that American orchestras conducting blind auditions hired more women. In most other cases though, only the first stage of the recruitment is made anonymous: this is the case in anonymous application procedures, such as the masking of identifying characteristics in resumes at the first selection stage.

In several European countries, pilot studies of the impact of such anonymization of resumes have been conducted, including relative large-scale field experiments in France, the Netherlands, Sweden and Germany. These experiments are summarized in **?**. Only a subset were truly randomized and we focus on discussion on this subset.[30]

The French government initiated an experiment in 2010 and 2011, 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 (**?**).

Among volunteer firms, resumes were either transmitted anonymously or non-anonymously. The experiments' main findings can be summarized as follows. First, women benefit from higher callback 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 callback 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 /citepduguet2010, which shows discrimination against minority candidates for some jobs, no discrimination for others, but never discrimination against majority candidates. /citet*behaghel 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% accepted the invitation. While participating were very

[30] **?** analyze an experiment conducted in parts of the local administration in 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 translate 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. The experiments focused on ethnic minorities. More specifically, a distinction is made between applicants with and without foreign (i.e., non-Western) sounding names. **?** emphasize in their study the lower callback rates for minority candidates with standard applications, but their analysis also reveals that these differences disappear 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 resumes are treated anonymously.

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 resumes). The anonymization therefore prevented selected firms from treating minority candidates more favorably during the experiment. Hence, 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 (**?**) (**?**).

Another large scale randomized field experiment took place in Germany in early 2010 (**?**). The publication of a correspondence testing study for Germany (**?**) triggered a lively public debate about discrimination in the hiring decisions of German firms.[31] 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. 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.[32] Unlike in the French study, the authors find that the anonymization leads to less discrimination against minority groups: anonymous hiring can reduce discrimination if discrimination is present beforehand (and can do the opposite if affirmative action is present beforehand). 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.

What would be nice to add to this literature is some evidence of the impact of anonymization on productivity. Not totally straightforward how to do that.

[31] 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.

[32] Four methods were considered: a) standardized application forms in which sensitive information is not included; b) refinements of existing online application forms such that sensitive information is disabled; c) copying applicant's non-sensitive information into another document; d) blackening sensitive information in the original application documents.

6 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 overarching 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 measurement of discrimination using the correspondence method. This body of work is truly remarkable in that it has demonstrated how 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 of these studies has added value. It is also one of these too rare cases where researchers have not shied away from replication (most likely because demonstrating differential treatment in their country was sufficiently important). On the other hand, researchers' ability to push the correspondence methods further to go beyond pure measurement of differential treatment has been more limited. Somewhat disappointingly, there has been close to no 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 that 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 **?** which demonstrate how correspondence studies can used to study the dynamics of discrimination (endogenous attention allocation in this case) suggest avenues for more creative uses. Maybe because so much economists' attention has been devoted to using field experiments to measure the extent of discrimination, there has been much less activity in designing creative field experiments to better document either the consequences of discrimination, or interventions that may 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) this field work could build upon. On the topic of consequence of discrimination, we are heartened to see a few recent papers such as **?** that develop a very creative field design to demonstrate how discrimination can be self-fulfilling.

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 been mainly centered around the contact hypothesis and exposure effects, while so many other strategies to de-bias have been proposed by psychologists and tested 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.

.