DE GRUYTER OLDENBOURG Analyse & Kritik 01/2016; 179–201

Marc Keuschnigg and Tobias Wolbring

The Use of Field Experiments to Study Mechanisms of Discrimination*

Abstract: This paper discusses social mechanisms of discrimination and reviews existing field experimental designs for their identification. We first explicate two social mechanisms proposed in the literature, animus-driven and statistical discrimination, to explain differential treatment based on ascriptive characteristics. We then present common approaches to study discrimination based on observational data and laboratory experiments, discuss their strengths and weaknesses, and elaborate why unobtrusive field experiments are a promising complement. However, apart from specific methodological challenges, well-established experimental designs fail to identify the mechanisms of discrimination. Consequently, we introduce a rapidly growing strand of research which actively intervenes in market activities varying costs and information for potential perpetrators to identify causal pathways of discrimination. We end with a summary of lessons learned and a discussion of challenges that lie ahead.

1. Introduction

The study of social inequalities, i.e., the unequal distribution of individual monetary and non-monetary resources within a population, has a long tradition in the social sciences. Inequalities can take various forms such as differences in employment and payment as well as unequal access to housing, education, and health care. It is common sense among social scientists that discrimination is a major social force in bringing about and perpetuating these differences (e.g., Quillian 2006; Reskin 2003). Discrimination, in this sense, refers to unequal treatment of individuals based on ascriptive characteristics such as age, ethnicity, and gender, which are irrelevant in the situation at hand. For example, in the labor market individuals might face disadvantages due to the color of their skin even though they are equally productive as whites. Similarly, certain ascriptive groups might be excluded from social exchange, although they are as trustworthy as others. Thus, our working definition of discrimination is more general than a narrow economic definition of discrimination which solely focuses on the distribution of material resources. Our understanding of the phenomenon is well in line with social psychological and sociological theories (e.g., Allport 1954; Goffman 1963;

* We thank Sonja Pointner, Merlin Schaeffer, the editors, and two anonymous reviewers for helpful comments. M.K. and T.W. contributed equally to this work.

Sherif 1958; Tajfel 1978), but also with economic theories of action (e.g., Becker 1990).

Common approaches to measure discrimination either rely on surveys of potential offenders' attitudes (see Krysan 2000 for racial attitudes) and potential victims' experiences (see Paradies 2006 for perceived racial discrimination) or on statistical analyses of social outcomes such as employment and wage (see Blinder 1973; Oaxaca 1973 for a prominent approach). Studies using survey designs claim to have provided ample evidence on the extent of discrimination and its change over time (e.g., Seibert/Solga 2005; see Kalter 2006 for a discussion). However, methodological drawbacks such as social desirability, subjective perceptions of maltreatment, and unobserved heterogeneity limit the rigor of these approaches (e.g., Bozoyan 2014). Avoiding these common pitfalls, unobtrusive field experiments became increasingly popular in recent years and have been applied in areas such as employment, housing, banking, and consumption (see Pager/Shepherd 2008; Riach/Rich 2002 for reviews). By now, several experimental protocols, particularly correspondence tests and in-person audits, have been well established and substantially extended our understanding of discrimination in various domains of social life. However, 'classical' experimental and quasi-experimental designs fail to identify social mechanisms behind the causal effect (Guryan/Charles 2013; Imai/Tingley/Yamamoto 2013). Hence, causal pathways which determine how heterogeneities transform into unequal life chances are yet poorly understood (see Diewald/Faist 2011 for a similar position).

In this paper we discuss social mechanisms of discrimination and review advanced research designs for their identification. First, we discuss two social mechanisms proposed in the literature to explain differential treatment based on ascriptive characteristics (*section 2*): Becker (1957), on the one hand, attributes discrimination to tastes for and against particular social groups. According to Becker's model the presence and extent of discrimination depends on the distribution of individual preferences, opportunities to discriminate without detection or sanctions, and—as will be explained later—the degree of market competition. Their interplay is the action-formation mechanism Becker proposes which jointly brings about the phenomenon of taste-based discrimination. Arrow (1973) and Phelps (1972), on the other hand, emphasize the role of stereotypes in reducing asymmetric information. Models of statistical discrimination draw on processes of information retrieval and belief formation to explain unequal treatment, while the situational mechanism also includes availability and reliability of information and opportunities for unequal treatment.

In *section 3* we present common approaches to study discrimination and discuss their methodological strengths and drawbacks. In *section 4* we elaborate why unobtrusive field experiments are a promising complement. For the dissection of mechanisms of discrimination, however, well-established experimental designs appear of limited use. Consequently, we introduce a rapidly growing strand of research which actively intervenes in market activities and systematically varies cost structures and information to identify the causal pathways of discrimination (*section 5*). We hope that reviewing this nascent field at its early stage of development inspires future sociological work. In the final section we

thus summarize lessons learned, among others, the importance of thorough theoretical reasoning to design experiments adequate for the isolation of social mechanisms and the role of analytical theory to overcome the context-dependency of field experiments conducted in real-world settings.

2. Mechanisms of Discrimination

Explanations of discrimination frequently refer to two distinct mechanisms guiding offenders' behavior: First, unequal treatment is proposed to stem from individual preferences where discriminators have tastes for and against interactions with certain ascriptive groups. Following Hedström's and Swedberg's (1998) typology, this is a micro-level action-formation mechanism. Second, statistical discrimination refers to decisions in situations of incomplete information where expectations of a counterpart's productivity, cooperativeness, or trustworthiness are inferred from easily observable characteristics (e.g., age, ethnicity, or gender). This process is a macro-micro situational mechanism (cf. Hedström/Swedberg 1998).[1]

2.1 Preference-based Discrimination

Gary Becker (1957) provided probably the best-known conceptualization of discrimination. He ascribes unequal treatment of individuals to affective tastes for and against particular social groups. This form of 'preference-based' discrimination is costly to the offender. Employers, for example, might offer higher wages to members of preferred social groups to avoid engagement of disliked—but no less productive—people. Consumers, on the other hand, might be willing to pay higher prices in order to evade personnel from a despised social group. In both cases discriminatory behavior comes at a monetary cost providing a (lower bound) estimate for the gravity of individual animus against certain ascriptive groups. In principle, this implication highlights that Becker's account on discrimination follows a 'wide' variant of rational choice theory: Offenders are willing to bear a price of discrimination precisely because objection against certain ascriptive groups is incorporated in their individual utility function. In other words,

[1] Although both formal models date back over 40 years in time we argue that they are well in line with current methodological demands brought forward by analytical sociologists: They are "abstract, precise, and action-based" (Hedström 2005, 37), have testable implications, and explicate (individual and structural) conditions which "regularly bring about a particular type of outcome" (Hedström 2005, 25). We thus agree with Kalter and Kroneberg (2014, 91) that analytical sociology offers "a new coat of whitewash" for an agenda that "has been at the core of methodological individualism, sociological rational choice theory, and explanatory sociology for some time". Nonetheless—as will become clear in the following—economic theories of discrimination are not fully satisfactory from the standpoint of analytical sociology. Due to their "instrumentalist attitude" (Hedström 2005, 37) the economics of discrimination lack specificity on the questions of where inputs (tastes, stereotypes) come from, which processes are set into motion, and how the micro and macro level interact. In addition, critics center on the lacking realism of theoretical assumptions and a deterministic model of man. Takács et al. 2014, for example, have proposed an agent-based model to overcome some of these deficits.

monetary costs of discriminatory behavior are offset by nonmonetary benefits for being spared from interaction with despised groups.

Taking into account the embeddedness of social exchange in market structures with different degrees of competitiveness, a second implication can be derived. Under conditions of fierce competition, in the long run, more efficient market actors drive less efficient rivals out of the market. The same does not necessarily hold for restricted markets in which, for example, segregation can occur. Thus, understanding the cost of unequal treatment as a tax levied on discriminators (Arrow 1973), the extent of discrimination should decline with rising market competition (see, however, Goldberg 1982). Firms, for example, should be less inclined to select job applicants from a particular group only if within-industry competition is fierce or supply of qualified employees is limited (see Blommaert/Coenders 2014). In the long run unequal treatment should thus disappear in contested markets as discriminators only survive in contexts of retrenched competition. Hence, tastes for discrimination interact with market structures in bringing about unequal treatment of ascriptive groups.

From the viewpoint of analytical sociology, however, Becker's account remains unsatisfactory as it solely reduces patterns of individual behavior to (collectively shared) preferences which are constant over time and context. It is left open to future inquiry which characteristics people actually discriminate on and, more importantly, how these preferences evolved and under which conditions they translate into unequal treatment. Consequently, one might warrant a theoretical explication of who bears discriminatory tastes and where these preferences originated from. In this context different schools of thought refer to historical processes molding human cognition and, among other things, tastes for other people. Without going into further detail these might be either long-term evolutionary developments (e.g., Buss 2005; Wilson 2000) bringing about preferences for in-group interactions or shorter-term social processes (e.g., Bourdieu 1984; Ridgeway 1991) shaping certain cultures of between-group interaction.

Interlinking the economic approach with social psychological theories of discrimination appears particularly promising for endogenizing preference formation. Theoretical work and experimental research on the minimal group paradigm (Tajfel 1978; Tajfel/Turner 1986) has shown when and how even arbitrarily assigned group markers lead to group conflict and discrimination. Akerlof and Kranton (2000; 2010) picked up this line of research, modifying Becker's approach by introducing the concept of identity into economics. Their key idea is that self-image can be threatened if others' presence or behavior contradicts one's own definition of the situation (e.g., females working in male occupations). Hence, the assignment of in-group and out-group status can change with the framing of the situation (e.g., male vs. female occupation). This situational variability of preferences then helps to explain why discrimination only occurs under specific conditions rather than unconditionally in all circumstances.

2.2 Statistical Discrimination

A competing explanation of discrimination emphasizes the role of cognitive stereotypes in reducing asymmetric information. Put forward by Kenneth Arrow (1973) and Edmund Phelps (1972) in the 1970s, this approach of 'statistical' discrimination assumes that in situations of incomplete information individuals enrich their assessments of other people with easily observable markers such as age, ethnicity, and gender. Based on these markers discriminators try to infer personal traits of interaction partners such as their productivity, cooperativeness, and trustworthiness. Expectations for the behavior of ascriptive groups can, for example, stem from previous interactions with persons sharing similar characteristics as well as from feedback received from external sources such as social networks, official statistics, and the media (Arrow 1998; Farmer/Terrell 1996). Employers, for example, might reject older job applicants based on the belief that—on average—older employees are less prepared to learn the ropes in a new working environment (Posner 1995). Similarly, female applicants of childbearing age may be turned down because an interruption of employment due to an upcoming pregnancy is more likely for this social group (Iversen/Rosenbluth 2010).

In contrast to unequal treatment grounded in animus, statistical discrimination can be individually profitable: As long as beliefs about an average characteristic within a particular social group are generally accurate, or are at least accurate for one's interaction partners, relying on inferences can indeed reduce asymmetric information for discriminators. This theoretical implication, however, is only true on average because beliefs about average characteristics ignore heterogeneities within ascriptive groups. Older job applicants who are prepared to incorporate new practices and, similarly, younger women who decide against motherhood deviate from their group's expectation value and are thus negatively discriminated. In contrast, people who underperform the ascribed group's average are positively discriminated. Although from a rational choice perspective one would assume that beliefs are formed rationally (e.g., on the grounds of Bayesian learning; Farmer/Terrell 1996) and should be at least approximately accurate (Aigner/Cain 1977), individuals exceeding or falling below their associated group's mean productivity, cooperativeness, or trustworthiness are discriminated by default.

Further, anticipation of statistical discrimination can trigger self-fulfilling processes among certain groups and thus contribute to socio-economic differences across society. Such 'self-fulfilling prophecies' (Merton 1948) occur if inaccurate expectations set processes into motion that finally validate initially false beliefs.[2] Thus, the potential for self-fulfilling prophecies relates inversely to belief accuracy (Jussim/Harber 2005). In contrast to Phelps' (1972) reliance on exogenous group differences in ability or signal reliability to model labor market

[2] The contrary case has been termed "inductively derived prophecy" (Biggs 2009) which describes a correct prediction of an outcome based on accurate expectations. While under inductively derived prophecies beliefs merely correlate with an outcome, beliefs are causal for an outcome under self-fulfilling prophecies.

inequalities, Arrow (1973) incorporated the self-fulfilling nature of inaccurate beliefs by endogenizing skill acquisition: Rational workers invest in their skills conditional on anticipated labor market chances (see Fang/Moro 2011 for formal details). Mere expectations of unequal payment can thus be sufficient to validate initially false beliefs.

Irrespective of whether initial beliefs are—on average—correct or incorrect, an important testable implication can be derived from statistical discrimination theories: Discrimination rooted in imperfect information should decline once information asymmetry is mitigated. If, for example, an employer learns about a worker's true characteristics she is less dependent on group-level inferences in assessing the worker's productivity. Farber and Gibbons (1996) as well as Altonji and Blank (1999) propose that, as distinguished from animus-driven discrimination, statistical discrimination declines with job seniority (see Altonji/Pierret 2001 for supportive empirical evidence). However, existing beliefs might be persistent retarding the decrease in discriminatory behavior (Pager/Karafin 2009). Also, the formation of group-level expectations is likely to be biased. This leaves room for animus-driven discrimination shaping human perception of how members of particular groups will behave. Hence, the conceptualizations of both preference-based and statistical discrimination can be intertwined posing serious challenges to empirical research trying to differentiate between these mechanisms. This is particularly true for the variety of formalizations and extensions of both models which often lead to similar predictions (see Fang/Moro 2011; Lang/Lehmann 2012 for reviews). Finally, as in Becker's model the concept of statistical discrimination provides no clear answer to the question of how individuals delineate social categories and which markers they use to classify others.

3. Empirical Approaches

Besides studies of law and legal records, common approaches to measure discrimination include surveys of potential offenders and victims, statistical analyses of social outcomes such as employment and wage, and experiments. In this section we highlight methodological strengths and weaknesses of these research strategies.

3.1 Surveys of Potential Victims and Offenders

The most intuitive approach to study discrimination is to ask targets of discrimination about their individual experiences. Surveys do not only allow the investigation of unequal treatment in structured situations of market exchange. Self-reported incidences of discrimination also offer a window into everyday social interactions. Hence, asking respondents about their subjective experiences is surely invaluable in providing a picture of the extent of perceived discrimination in various domains of social life. At least since the classical work by Thomas and Thomas (1928, 571–572) the sociological relevance of subjective perceptions is widely acknowledged: "If men define situations as real, they are real in their

consequences." This proposition similarly holds for unequal treatment on the basis of ascriptive characteristics: Krieger (2014), for example, highlights that reported experiences of discrimination are correlated with other psychological (e.g., anxiety, well-being) as well as physiological measures (e.g., blood pressure, diseases).

Surveys of victims can also provide suggestive evidence on potential causal pathways of unequal treatment. Job applicants, for example, can give detailed information on structural constraints during the application process as well as on the content and procedure of job interviews. This data might help researchers to shed light on both obvious and subtle forms of discrimination which are hard to measure otherwise (e.g., increasing demands for the job, asking of overly difficult questions, creating an uncomfortable interview situation).

However, surveys of victims are inadequate to measure the actual intensity of discrimination: "What remains unclear from this line of research [...] is to what extent perceptions of discrimination correspond to some reliable depiction of reality." (Pager/Shepherd 2008, 183) Several arguments corroborate these doubts: First, potential victims often lack information necessary to identify a behavior as discriminatory. For example, applicants can hardly tell whether they were treated fairly without knowing the applicant pool, the full list of job demands, or the continuum of conditions (e.g., salaries, working hours) from which the employer can choose her offer. Second, in these surveys the presence or absence of discrimination lies in the eye of the beholder. Still, people differ in their evaluation of acceptable behavior and thus it remains unclear whether group differences result from differences in interpretation or differences in treatment. Third, social desirability might bias estimates. Underreporting, for example, occurs in the case of sensitive topics such as sexual harassment, while overestimation can result from ex post rationalizations for negative events and sensitization due to public campaigns.

Due to these limitations it appears natural to scrutinize attitudes, stereotypes, and behavior of potential offenders. Probably the most widely applied line of research in this area is the study of attitudes and stereotypes which has documented that ascriptive characteristics such as age, ethnicity, and gender regularly activate general expectations about the personality and behavior of a person (Ridgeway 1991). Furthermore, as Midtbøen and Rogstad (2012) argue, in-depth interviews with employers might help to dissect the causes of discrimination in the labor market by providing detailed information on hiring strategies and firms' current economic condition. However, surveys of potential perpetrators face rather similar problems as surveys of potential victims. Social desirability is obviously a pressing issue as are ex post rationalizations for behavior. Particularly in the case of offences with potential legal consequences it would be naive to expect perpetrators to be honest and admit misbehavior. In addition, stereotypes do not directly translate into behavior (LaPiere 1934; Pager/Quillian 2005). As Petersen (2009) and Reskin (2003) emphasize, for a given set of stereotypes the extent of discrimination varies substantially with structural conditions such as the organizational and institutional setting.

3.2 Statistical Analysis of Social Outcomes

Due to the inherent difficulties of directly measuring discrimination on the basis of subjective reports from targets and perpetrators many researchers stick to an indirect identification strategy using observational data (e.g., Blinder 1973; Oaxaca 1973). The main idea of this line of inquiry is to ask for gaps in social outcomes, such as employment and wage, between ascriptive groups after partialling out differences in their social-structural composition. Usually, all available variables that might be associated with the treatment or outcome of interest are adjusted for. This so-called residual approach is widely applied particularly in studies of labor market discrimination (see Altonji/Blank 1999 for a review): In this literature researchers regularly control for standard measures of education and work experience as suggested by human capital theory (e.g., Mincer 1974) and, if available in the data, for further covariates, such as intelligence, social competence, and mental and physical health. The remaining effect of the ascriptive characteristic on employment status or wage is then attributed to discrimination.

Other authors have heavily criticized this identification strategy calling it a "shotgun" (Lieberson 1985, 39) or "kitchen sink" approach (Wooldridge 2004, 4). One concern is the possibility of overcontrol. Researchers do not explicate the theoretical assumptions about the causal relationships between the covariates in the model and might, therefore, falsely control for mediating or collider variables (see Bozoyan/Wolbring 2015; Elwert/Winship 2014; Winship/Morgan 2015 for details). Another concern is omitted variable bias. Using observational data one cannot test for and, hence, can never fully rule out the possibility that unobserved differences in productivity between the groups under investigation confound estimates of the extent of discrimination. The same holds for other forms of unobserved heterogeneity (e.g., unmeasured differences between firms, occupations, and positions; see Petersen 2009) which undermine the conditional independence assumption. Obviously, regression-based approaches share this problem with other research strategies such as matching and decomposition methods.[3]

3.3 Experimental Approaches

The discussion of observational approaches has highlighted two important issues: First, subjective reports of experiences and behavior are too unreliable to measure the actual extent of discrimination. Second, without exogenous variation of the 'ascriptive group' treatment it is nearly impossible to provide definite empirical evidence for the presence and intensity of discrimination. Laboratory experiments seem to solve both problems. On the one hand, subjects usually make costly decisions in the lab providing various outcome indicators as objective measures of discrimination. On the other hand, random allocation of

[3] It is well worth mentioning that fixed effects regression (Allison 2009; Brüderl/Ludwig 2015), a powerful method to control for time-constant unobserved heterogeneity, is usually not applicable to the study of discrimination, because many ascriptive characteristics of interest are constant over time.

treatments, which are under the control of the researcher, secures the conditional independence assumption for the stimulus and thus attenuates concerns about causality and biased estimates. These and further advantages of experimental methodology have also been recognized by analytical sociologists (e.g., Bohnet 2009). Although controlled experimental interventions do not automatically permit identification of causal mechanisms (see Imai/Tingley/Yamamoto 2013), an interventionist approach is well in line with analytical sociologists' notion of causation (see Hedström/Ylikoski 2010) and has a solid theoretical foundation in the counterfactual framework of causality (see Morgan/Winship 2015 for an introduction).[4]

Clearly, as for example Heckman (2005) as well as Charles and Guryan (2011) argue, it is impossible to systematically vary a person's actual gender, race, or height. Furthermore, other characteristics of interest, such as age, looks, and weight, can only be changed to a certain degree. In this sense, experimental research is limited by the inability to construct a perfect counterfactual. However, as Sobel (2005) points out in his rejoinder to Heckman (2005), we can vary people's perceptions of these characteristics while holding everything else constant (see also Imbens/Rubin 2015). For example, by now many laboratory experiments use names, places of birth or living, and pictures to signal ascriptive group membership.

Although laboratory experiments have clear advantages for the endeavor of causal inference (see Angrist/Pischke 2009; Morgan/Winship 2015), many fear that the external validity is severely limited making them nearly useless for most social science research questions. This argument, however, is less convincing than it might appear at first glance: First, one is driven to ask what it means to generalize empirical results which are not internally valid. According to Morton and Williams (2010, 385) it does not make sense to establish the external validity for results that have not been demonstrated to be internally valid. In the same line Falk and Heckman (2009) argue that the foremost task of science is to get the causal effect right (see also Campbell 1957). Second, the question is not whether a finding is generalizable, but to which social settings it is transportable (again see Falk/Heckman 2009). In this sense every finding is externally valid, but only for a certain set of conditions. Third, internal and external validity are not properties of a design but of inferences (Shadish/Cook/Campbell 2001, 34). Hence, the degree of external validity of inferences drawn from lab experiments depends on the content of the research and especially on the sensitivity of the topic under study. For example, studies using both lab and field experiments on prosocial behavior revealed a substantial degree of accordance (Franzen/Pointner 2013), but validation studies for many other topics are sorely needed and open to future research (see Camerer 2015; Levitt/List 2007 for a current discussion in experimental economics).

---

[4] The concept of mechanisms has somewhat different meanings in both schools of thought. While analytical sociologists define *social mechanisms* as a constellation of entities and activities that regularly bring about a phenomenon (Hedström 2005), the concept of *causal mechanisms* seems to be more specific as it implies the presence of a mediating variable ($X \rightarrow Z \rightarrow Y$; Imai et al. 2011).

Nonetheless, doubts about the external validity of laboratory experiments have inhibited their application in the social sciences, especially in sociology, and have led to an increased interest in field experimental research (see Jackson/Cox 2013). While field experiments share major strengths with lab designs, they are not staged in the artificial and potentially reactive laboratory setting but conducted under 'natural' conditions (Gerber/Green 2012; Levitt/List 2008). We thereby recur to Cronbach's (1982) concept of *utos* to emphasize that naturalness encompasses not only the experimental setting (s), but also the units (u), treatments (t), and observing operations (o) (see Wolbring/Keuschnigg 2015 for details). Thus, naturalness is not a dichotomous characteristic but encompasses different dimensions and degrees. Clearly, a high degree of naturalness in regard to all four dimensions of experimental design helps to attenuate concerns about external validity. Furthermore, social desirability bias and reactivity can be substantially reduced in field experiments as compared to lab experiments if unobtrusive measures of behavior—such as incidents of aggression (Cohen et al. 1996), prosocial behavior (Milgram/Mann/Harter 1965), and littering (Keizer/Lindenberg/Steg 2008)—are employed. Obviously, due to the danger of social desirability bias the use of such measures is particularly important in the context of legally sanctioned behavior, as is the case for many forms of discrimination in Western societies.

In the remainder of this article we discuss what field experiments can additionally contribute to dissect the anatomy of discrimination. In section 4 we review widely applied variants of unobtrusive field experiments in research on discrimination. Section 5 then presents new and more sophisticated designs that do not only seek to estimate the extent of unequal treatment but try to separate specific mechanisms of discrimination.

4. Audit Designs

Pioneered by British sociologists in the 1960s and 1970s (Daniel 1968; Jowell/Prescott-Clarke 1970) the use of unobtrusive field experiments has a long tradition in the study of discrimination (see also LaPiere 1934 for a much earlier application). Accordingly, experimenters can draw on a range of well-developed designs to estimate the extent of discrimination in various situations of social selection. Labeled as audit studies these protocols include fictitious letters of application (correspondence tests) and in vivo interviews (in-person audits). We briefly discuss these protocols with regard to discrimination in the labor market (see Pager/Shepherd 2008; Riach/Rich 2002 for reviews).

Correspondence tests (see Bertrand/Mullainathan 2004 for a widely cited application) consist of two or more fake application letters sent to a sample of employers in response to job postings. Applicants' résumés signal equal productivity, however, vary at least in one dimension reflecting ascriptive differences across job-seekers (e.g., in age, ethnicity, or gender). Holding other factors constant this design aims at identifying the causal effect of ascriptive features on response rates (usually by telephone or email). Most importantly, the design

permits elicitation of discriminatory tendencies using direct, unobtrusive measurement of potential offenders' behavior.

Given its simplicity the correspondence test comes with a few weaknesses. Suspending an unwanted limitation to written applications, in-person audits offer valuable methodical benefits. First, physical participation allows collection of additional response variables and detailed information on the job interview. Second, when performed well, human testers can increase treatment reliability, as—unlike in correspondence tests where experimenters remain unsure whether the potential offender actually perceives the stimulus—stimulus dosage is under more direct control. As we will see below, however, others argue that particularly this aspect of in-person audits poses methodological problems seriously threatening treatment integrity and causal inference: In contrast to correspondence tests, in-person audits lack randomization of treatments. Instead, they regress to a quasi-experimental design as testers themselves bring along the ascriptive characteristic of interest.

In a seminal implementation of an in-person audit, Pager, Western, and Bonikowski (2009) sent ten young men in rotating teams of three (white, Latino, black) to take part in 340 interviews for blue-collar jobs in New York. Testers were selected from more than 300 applicants to match in physical appearance and style of communication. For each interview testers were randomly equipped with one of three substantially equivalent résumés. Ultimately, 31 percent of white, 25 percent of Latino, and 15 percent of black confederates received a callback. Even after imposing a trade-off between skin color and résumé (white testers received the stigma of a 18 month prison sentence for a drug felony) response rates for whites (17 percent) slightly exceeded callbacks for Latinos (15 percent) and blacks (13 percent). Debriefings following each fielding permitted collection of additional outcome variables. Most importantly, these included qualitative information on perceived responses to stimuli and revealed frequent redirection of minority applicants to jobs other than the one which had been originally posted. Particularly for black candidates this included positions with lower salary, less customer contact, and minor professional status.

For a causal interpretation of these results successful matching of testers is a basic requirement. Due to their personal participation in the experimental interaction, experimenters must synchronize far more personal characteristics than is required in written letter designs. Also, researchers can hardly control the experimental situation as is the case in correspondence tests. This is particularly true for job interviews where the course of conversation cannot be fully scheduled and requires some degree of improvisation by the tester. Apparently, it is impossible to control for all potentially relevant characteristics offenders might use to discriminate on.[5] Due to the lack of randomized stimuli in-person audits are threatened by unobserved heterogeneity (Heckman/Siegelman 1993). Fol-

[5] The set of personal features which need to be controlled for varies with the social situation under observation. Matching is relatively straightforward for short-time interactions (e.g., product purchases, taxi rides) and gets increasingly difficult with rising complexity of the experimental situation. Telephone audits offer a valuable middle ground permitting multidimensional stimuli (e.g., names, dialects) but rendering physical matching unnecessary (cf. Riach/Rich 2002). A popular telephone design frequently used in the study of discrimination is

lowing Pager, Western, and Bonikowski (2009), however, this general weakness of experimental interviews can be alleviated by thorough training and supervision of testers, a rotating composition of tester teams, and ex post facto control for potential experimenter effects. Statistical control for tester effects is particularly important also because in-person audits do not use a double blind protocol (Heckman/Siegelman 1993). Instead, testers are typically informed about the research question and might jeopardize treatment validity, for example, by causing rejections due to subjective anticipation of unequal treatment.

Then again, successful matching can—both in correspondence tests and in in-person audits—bias estimates of discriminatory behavior. Most prominently, Heckman (1998) argues that strict comparability of résumés and testers forces potential offenders to arbitrary decisions for or against certain applicants. Consequently, audit designs should be expected to over-estimate discriminatory behavior as, for example, employers overrate the relevance of ascriptive characteristics to break ties between otherwise equivalent job-seekers. To limit over-estimation most studies use low-threshold response variables. In the case of hiring, for example, callbacks are clearly less exclusive than binding job offers. Attention should also be paid to generally low response rates in employment studies, limiting the statistical power of the audit design. This is aggravated by the fact that some employers follow a strict response policy calling back each applicant regardless of her chances for employment. In this case, potential discrimination can only be estimated from data provided by employers without such a callback policy.

Naturally, these field experiments also pose serious ethical concerns, particularly in the case of labor market implementations (see also Riach/Rich 2004 for a discussion). First, participants are deceived in a sensible and institutionally highly regulated domain. Appropriate anonymization is thus an absolute requirement. Second, attention should be paid to participants' effort for screening fictitious applicants. Typically, audits thus focus on entry-level job postings to limit screening time. Third, researchers should be aware of potential negative externalities for third party applicants. These might arise, for example, if a real candidate gets rejected in favor of a fictitious applicant. To restrict unintended consequences researchers must decline positive responses to fake applications immediately.[6]

---

the 'wrong number technique' (Gaertner/Brinkmann 1971), in which testers pretend misdialing and ask the receiver for a favor (e.g., calling a mechanic to fix a broken-down car).

[6] Factorial survey experiments circumvent these ethical concerns and allow location of discriminatory tendencies in both large survey samples and specific populations of potential offenders (see Auspurg/Hinz 2015 for an overview). Typically, survey respondents are confronted with brief descriptions (vignettes) of hypothetical situations or persons featuring specific attributes (dimensions). In labor market applications (e.g., Jasso/Webster Jr. 1999; Williams/Ceci 2015) respondents are asked which fictitious job applicant they would be willing to hire. Similar to the correspondence test ascriptive dimensions are randomized across vignettes and respondents securing causal interpretation and high internal validity. Detriments may arise from reactivity such that researchers should be aware of social desirability potentially biasing results (Auspurg et al. 2015) and limited correlation of (measured) behavioral intentions and (unmeasured) actual behavior (Hainmueller/Hangartner/Yamamoto 2015).

The portrayed field designs provide estimates of discriminatory tendencies in a variety of real-world social settings. Because discrimination within these 'classical' set-ups occurs in spite of equivalent résumés, proponents of the audit method frequently refer to the preference-based approach (Becker 1957) to explain unequal treatment across social groups. This interpretation, however, is only valid if one accepts that confounders (i.e., all potentially relevant personal features of fictitious applicants) are fully controlled for and thus unobserved heterogeneity does not threaten internal validity (Heckman/Siegelman 1993). Indeed, careful manipulation of résumés in order to impose trade-offs between, for example, individual productivity and ascriptive characteristics can provide some insights into the nature of discrimination. Even if extensive résumés are provided, however, one cannot safely reject that participants infer some additional information from ascriptive features and thus discriminate statistically (Arrow 1973; Phelps 1972). As Heckman (1998) and Neumark (2012) point out, differences in employment and payment can even occur if groups' mean productivity is similar but perceived variances or precision of available signals differ between groups. For example, even though résumés are equivalent a potential employer might expect more variation in the (unobserved) abilities of blacks than of whites. Hence, we conclude that neither correspondence tests nor in-person audits allow for an identification of social mechanisms behind discriminatory behavior. In the following we introduce a new class of field experimental designs which may help to shed further light on the causal pathways.

5. Extended Field Experimental Designs

Attempts of separating mechanisms of discrimination by experimental design have to meet practical demands way beyond the provision of integer treatments. Such set-ups require a much larger degree of control over the experimental situation as they typically combine an audit design with active interventions into market activity. Using a variety of designs where subjects decide on discriminatory behavior in the face of varying costs and information, a nascent literature sheds light on the causal pathways of discrimination (see List 2004; Zussman 2013 for further applications). Methodological advances have not been fully convincing yet as studies merely present suggestive evidence on the mechanisms governing discriminatory behavior. As a starting point for more sophisticated—and sociologically motivated—designs in the future, however, we discuss two exemplary field experiments which at least give indirect indication on the causes of discrimination both at the workplace and in consumer markets.

Using the potentials of the internet for field experimental research,[7] Doleac and Stein (2013) sold iPods on local online platforms varying, among other things, ethnicity and social status of fictitious sellers. Controlling for unobserv-

[7] Unobtrusive online studies are relatively easy to implement even at large scale and typically offer a large degree of control over the experimental situation (see Blommaert/Coenders 2014 for a recent application). Although they particularly raise new ethical questions and practical challenges, compared to other field experiments they can be designed to be less prone to reactivity and other distortions to causal inference (Golder/Macy 2014).

able features, ascriptive characteristics were signaled using photographed hands presenting the item for sale. Hands were either black or white (indicating ethnicity) while some white hands sported a clearly visible wrist tattoo (indicating low social status). 1,200 advertisements for an iPod Nano were successively posted in more than 300 geographically focused online markets. In contrast to large platforms such as eBay and Amazon, these local markets lacked formal systems of bidding, completion, and reputation leaving it to anonymous buyers and sellers to directly negotiate price and delivery by email. The markets' local focus minimized the risk of cross-contamination between treatment conditions (satisfying the so-called stable unit treatment value assumption)[8] and facilitated parallel intervention under varying market conditions.

Regarding both ethnicity and social status the experimental effects revealed a considerable degree of discrimination among buyers: As compared to whites, advertisements by black sellers were twice as likely to be removed prematurely, received 13 percent fewer and less trustworthy responses and generated 12 percent lower prices. Tattooed sellers similarly underperformed 'non-suspicious' whites although, at least on some response variables, to a fairly smaller extent than blacks. Additional analyses then exploited differences between local markets to indirectly test for mechanisms of discrimination. This quasi-experimental study of interaction effects showed that—consistent with Becker's implication for preference-based discrimination—price gaps between whites and blacks as well as between 'non-suspicious' and tattooed sellers diminished with intensified competition among buyers. More importantly, however, black sellers received lowest prices in areas of high property crime and intense racial segregation whereas prices proposed to tattooed whites were largely unaffected by race-related disparities across regions. Although limited both in statistical precision and causal interpretability, the interactions reported are consistent with anxious buyers using ethnicity as an indication of sellers' good will. Also, the experimental advertisements did not specify delivery modes (sold iPods were eventually shipped by mail). As delivery in local online markets frequently includes collection at residence, buyers might shun a trip to a poor and potentially dangerous neighborhood associated with black sellers. From these results Doleac and Stein (2013, F489) suggest that rather than "indulging in taste-based discrimination" buyers used skin color as a proxy for unobservable characteristics to statistically discriminate between sellers.

Despite its obvious achievements the study can also be criticized on methodological grounds. First, the authors introduced tattoos as an indicator of low social status (see Gambetta 2009 for tattoos' signaling properties). Although their results clearly show a causal effect of tattoos on several outcome measures,

[8] Rubin 1974; 2008 posits the stable unit value assumption (SUTVA) as one of the fundamental preconditions for counterfactual causal inference. In a nutshell, the SUTVA holds if there is "no interference between units [...] leading to different outcomes depending on the treatments other units received and there are no versions of treatments leading to 'technical errors"' (Rubin 1980, 591). Thereby, 'technical errors' originate from the fact that we can only conduct a finite number of replications of an experiment which can lead to differences between 'true' potential outcomes and average observable outcomes due to varying ancillary experimental conditions (see Neyman/Iwaszkiewicz/Kolodziejczyk 1935).

treatment validity remains obscure. Whether buyers indeed associate tattoos with low social status is not directly tested and might be doubted. For example, some indicators previously suited to signal social status might have changed their meaning over time (see, e.g., Schiermer 2014 on tattoos in 'hipster culture'). To rule out these concerns manipulation checks could be easily implemented by presenting the photographs to several independent evaluators asking them to rate expected attributes of fictitious sellers.

Second, and more importantly, refinements of the suggested set-up should clearly seek to measure statistical discrimination by design rather than by quasi-experimental proxies. A possible implementation could include the use of pre-installed reputation systems which have been shown to ease anonymous online transactions and improve highly ranked sellers' market outcomes by providing reliable information on individual trustworthiness and reciprocity (e.g., Diekmann et al. 2014). Assuming that statistical discrimination is a helpful strategy to assess interaction partners' hidden characteristics under imperfect information (Arrow 1973; Phelps 1972), ascriptive markers should cease to affect market outcomes once information asymmetry is mitigated. Any remaining inequalities in individual outcomes could then be interpreted as stemming from animus-driven discrimination. A careful—and admittedly costly—manipulation of fictitious sellers' reputation which gradually improves potential buyers' information would clearly strengthen the design's relevance as a blueprint for further experimentation.

Our second exemplary study in parts follows this identification strategy. Employing 169 Danish school students (aged 16–20) for a large mailing task, Hedegaard and Tyran (2014) used a fully controlled intervention to manipulate decision makers' costs and information. Recruited to prepare letters at piece rate for 2×90 minutes, students were first invited to work single-handedly in separate rooms at the University of Copenhagen. This first stage of the experiment provided an individual measure of students' productivity. For a second round, students were informed that letters would be prepared in teams of two and piece rate payments would be shared with the coworker. Randomly selected students were then asked to choose a coworker from two possible candidates. At this point the authors used existing differences in students' first names—signaling either a Danish or Muslim family background—to systematically vary ethnicity within the pools of proposed partners. Candidates had the same gender as the focal subject but came from different schools. To minimize social desirability the selection of coworkers was tied to a choice between two possible work days (each participant indicated full availability at recruitment).

In the first treatment condition statistical discrimination was eliminated by design, providing decision makers with full information both on candidates' ethnicity and first round productivity. As candidates' productivity differed, many subjects were confronted with a trade-off between choosing a same-type coworker and maximizing expected earnings. In contrast to the audit design, decision makers thus had to pay a varying price for discrimination (6.7 Euros on average) permitting a direct test for preference-based discrimination. As a result, 38 percent of participants engaged in animus-driven discrimination, on average

renouncing 8 percent (5 Euros) of round two earnings to work with a same-type partner. Moreover, there was clear support for Becker's hypothesis of market competition: Increasing the price for unequal treatment by 10 percent reduced discriminatory choices of coworkers by 9 percent. Revealing only candidates' ethnicity, a second treatment condition allowed for both forms of discrimination. As Danes were in fact more productive than participants with Muslim-sounding names (116 vs. 100 letters packed in the first round) statistical discriminators should be expected to solely choose Danish coworkers. Despite income losses, however, especially Muslim participants opted for non-Danish coworkers. A necessary condition to reject statistical discrimination is that heterogeneous teams are as productive as homogeneous teams. This was the case such that this alternative explanation of behavior could be ruled out.

In an additional attempt to disentangle proper statistical discrimination from potentially biased stereotypes, Hedegaard and Tyran pulled up another sample of Copenhagen students surveyed to provide (incentivized) beliefs about type-specific productivities in the letter-packing task. Indeed, elicited beliefs somewhat mirrored actual productivity differences disclosed in the field experiment providing indirect indication of fairly accurate group level expectations among participants. Hence, the authors conclude that—although some of the price paid by discriminators was indeed due to biased stereotypes—team building decisions were taken under a considerable degree of preference-based discrimination.

The design is well suited to measure individual willingness to pay for discrimination. Apparently, the authors did not settle with rejecting statistical discrimination as purposeful behavior to reduce asymmetric information but tried to disentangle 'accurate' statistical discrimination from biased stereotypes. The rejection of strongly biased beliefs about ethnicities' average productivity, however, relied on external elicitation among similar yet non-participating students. Once again one would prefer direct identification by design. First steps could be made by asking a subgroup of decision makers about their expected group output in round two given they worked with either candidate. Also, future researchers can draw on a range of incentivized and potentially less intrusive belief elicitation methods (e.g., Trautmann/van de Kuilen 2014) such as the possibility of placing bets on candidates' expected productivity.

Comprising both studies an interpretation could read as follows: Statistical discrimination seems to dominate in social settings of imperfect information, limited interaction, and required trust. Animus, on the other hand, inevitably influences behavior when unequal treatment is inexpensive. Hence, both preference-based and statistical discrimination can be intertwined exacerbating empirical identification of separate pathways of discrimination. However, a final note of caution should be sounded: The proposed designs share problems with correspondence tests in general (see, again, Heckman 1998). The researcher has to rely on proxies such as names, color of the skin, and tattoos to signal the theoretical constructs of interest. The assumption that these treatments solely influence subjects' perception of the intended construct is often problematic. For example, a name may not only indicate ethnicity, but also status and social origin. This

can undermine treatment validity and may result in misleading results regarding the ascriptive characteristic that actually causes unequal treatment.

6. Conclusion

In this paper we discussed social mechanisms of discrimination and proposed advanced research designs for their identification. As a starting point we reviewed two social mechanisms proposed in the literature—animus-driven and statistical discrimination—to explain differential treatment based on ascriptive characteristics. We then presented common approaches to study discrimination, discussed their methodological strengths and drawbacks, and elaborated why unobtrusive field experiments are a promising complement. Since well-established field experimental designs often fail to identify the mechanisms of discrimination, we introduced a rapidly growing strand of research which actively intervenes in market activities and systematically varies cost structures and information to identify the causal pathways of discrimination.

In this concluding paragraph we want to highlight three lessons learned from the previous discussion and emphasize theoretical and methodological challenges that lie ahead. First, thorough theoretical reasoning is essential to ensure both internal and external validity of field experiments. Clear ex ante prediction about causal pathways are indispensable to design experiments adequate for the isolation of social mechanisms. This concerns the measurement of process variables, assumptions about their relationships with treatments, outcomes, and other mediators, as well as empirical strategies for the identification of indirect causal effects (for discussions see Card/DellaVigna/Malmendier 2011; Imai/Tingley/Yamamoto 2013; Ludwig/Kling/Mullainathan 2011). In discrimination research, where causal pathways intertwine, the separation of mechanisms is a particularly demanding challenge for experimental protocols. Analytical theory also plays a crucial role to overcome the context-dependency of field experiments conducted in real-world situations. Since every study is based on specific units, treatments, observing operations, and social settings, it remains unclear whether empirical findings can be transported to different conditions (Falk/Heckman 2009). At this point theory contributes by specifying the conditions under which a particular treatment should be effective (Deaton 2010). Thus, a stronger interlinkage of field experiments and social theory can help to mitigate concerns about generalizability, particularly if small-scale experiments are supposed to offer answers to 'big' questions (e.g., the effect of institutions on discrimination).

Second, the search for mechanisms is not just worthwhile due to academic curiosity. Instead, identifying the deeper roots of discrimination is a crucial requirement for the design of effective interventions to prevent and reduce unequal treatment (Reskin 2003). Practical countermeasures can be applied at the organizational level of jurisdiction, unions, and firms, where specific interventions must be fine-tuned to the specific type of discrimination present (e.g., tolerance campaigns against prejudices, information campaigns against biased

stereotypes). In addition, knowledge about the pathways of unequal treatment also enables victims to countersteer discriminatory tendencies (e.g., avoiding stereotypical behavior, signaling competence and motivation). Certainly, the extent of specific forms of discrimination depends on boundary conditions, particularly the social situation at hand. To sufficiently account for this context dependency, future extensions to a variety of situations of social selection are sorely needed.

Finally, the renaissance of field experiments in the social sciences, particularly the case of online field experiments (see Golder/Macy 2014; Wolbring/Keuschnigg 2015), also implies new concerns and pitfalls. Foremost, new ethical issues and questions about data privacy arise. For example, a recent online experiment on *Facebook* by Kramer, Guillory, and Hancock (2014) initiated a heated debate as to the extent to which research should and may intervene in online settings, manipulate the appearance of websites, and collect data. The discussions focus on the violation of personal rights, ethical and practical problems of deception, negative effects of experimental stimuli, exclusion of subjects from effective treatments, goal conflicts, unintended consequences and other momenta of field work, and field access. These different aspects have to be weighed up against each other and against the scientific and practical importance of research findings. Surely, the dignity and well-being of participants is inviolable. However, as Riach and Rich (2004) argue, uncovering practices of discrimination and understanding perpetrators' motivations is also a highly valuable public good which should not be given up too easily for other important, but less urgent reasons such as informed consent and avoidance of deception. This is especially true if minimal harm and inconvenience are inflicted on subjects under study, and if reported results are anonymized by using highly aggregated statistical indicators.

Bibliography

Aigner, D. J./G. Cain (1977), Statistical Theories of Discrimination in Labor Markets, in: *Industrial and Labor Relations Review* 30, 749–776

Akerlof, G./R. Kranton (2000), Economics and Identity, in: *Quarterly Journal of Economics* 115, 715–753

— (2010), *Identity Economics: How Our Identities Affect Our Work, Wages, and Well-being*, Princeton

Allison, P. D. (2009), *Fixed Effects Regression Models*, Thousand Oaks

Allport, G. W. (1954), *The Nature of Prejudice*, Garden City

Altonji, J. G./R. M. Blank (1999), Race and Gender in the Labor Market, in: Ashenfelter, O./D. Card (eds.), *Handbook of Labor Economics, Volume 3C*, Amsterdam, 3143–3259

—/C. R. Pierret (2001), Employer Learning and Statistical Discrimination, in: *Quarterly Journal of Economics* 116, 313–350

Angrist, J./J.-S. Pischke (2009), *Mostly Harmless Econometrics. An Empiricist's Companion*, Princeton

Arrow, K. J. (1973), The Theory of Discrimination, in: Ashenfelter, O./A. Rees (eds.), *Discrimination in Labor Markets*, Princeton, 3–33

— (1998), What Has Economics to Say About Racial Discrimination?, in: *Journal of Economic Perspectives* 12, 91–100

Auspurg, K./T. Hinz (2015), *Factorial Survey Experiments*, Thousand Oaks

—/—/C. Sauer/S. Liebig (2015), The Factorial Survey as a Method for Measuring Sensitive Issues, in: Engel, U./B. Jann/P. Lynn/A. Scherpenzeel/P. Sturgis (eds.), *Improving Survey Methods: Lessons from Recent Research*, New York, 137–149

Becker, G. S. (1957), *The Economics of Discrimination*, Chicago

— (1990), *The Economic Approach to Human Behavior*, Chicago

Bertrand, M./S. Mullainathan (2004), Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination, in: *American Economic Review* 94, 991–1013

Biggs, M. (2009), Self-Fulfilling Prophecies, in: Hedström, P./P. Bearman (eds.), *The Oxford Handbook of Analytical Sociology*, Oxford, 294–314

Blinder, A. S. (1973), Wage Discimination: Reduced Form and Structural Estimates, in: *Journal of Human Resources* 8, 436–455

Blommaert, L./M. Coenders (2014), Discrimination of Arabic-Named Applicants in the Netherlands: An Internet-Based Field Experiment Examining Different Phases in Online Recruitment Procedures, in: *Social Forces* 92, 957–982

Bohnet, I. (2009), Experiments, in: Hedström, P./P. Bearman (eds.), *The Oxford Handbook of Analytical Sociology*, Oxford, 639–665

Bourdieu, P. (1984), *Distinction: A Social Critique of the Judgement of Taste*, Cambridge

Bozoyan, C. (2014), *Schwer im Nachteil: Zur Diskriminierung übergewichtiger und adipöser Menschen in Schule und Arbeitsmarkt*, Hamburg

—/T. Wolbring (2015), The Usefulness of DAGs: What Can Directed Acyclic Graphs Contribute to a Residual Approach to Weight-related Income Discrimination?, in: *Journal of Applied Social Science Studies* 135, 1–14

Brüderl, J./V. Ludwig (2015), Fixed-Effects Panel Regression, in: Best, H./C. Wolf (eds.), *Regression Analysis and Causal Inference*, Thousand Oaks, 327–357

Buss, D. M. (2005) (ed.), *Handbook of Evolutionary Psychology*, Hoboken

Camerer, C. F. (2015), The Promise and Success of Lab-field Generalizability in Experimental Economics: A Reply to Levitt and List, in: Frechette, G./A. Schotter (eds.), *The Methods of Modern Experimental Economics*, Oxford, 249–295

Campbell, D. T. (1957), Factors Relevant to the Validity of Experiments in Social Setting, in: *Psychological Bulletin* 54, 297–312

Card, D./S. DellaVigna/U. Malmendier (2011), The Role of Theory in Field Experiments, in: *Journal of Economic Perspectives* 25, 39–62

Charles, K./J. Guryan (2011), Studying Discrimination: Fundamental Challenges and Recent Progress, in: *Annual Review of Economics* 3, 479–511

Cohen, D./R. Nisbett/B. Bowdle/N. Schwarz (1996), Insult, Aggression, and the Southern Culture of Honor: An 'Experimental Ethnography', in: *Journal of Personality and Social Psychology* 70, 945–960

Cronbach, L. J. (1982), *Designing Evaluations of Educational and Social Programs*, San Francisco

Daniel, W. W. (1968), *Racial Discrimination in England*, Harmondsworth

Deaton, A. (2010), Instruments, Randomization, and Learning about Development, in: *Journal of Economic Literature* 48, 424–455

Diekmann A./B. Jann/W. Przepiorka/S. Wehrli (2014), Reputation Formation and the Evolution of Cooperation in Anonymous Online Markets, in: *American Sociological Review* 79, 65–85

Diewald, M./T. Faist (2011), From Heterogeneities to Inequalities: Looking at Social Mechanisms as an Explanatory Approach to the Generation of Social Inequalities, in: *SFB 882 Working Paper Series* 1, Bielefeld

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

Elwert, F./C. Winship (2014), Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable, in: *Annual Review of Sociology* 40, 31–53

Falk, A./J. J. Heckman (2009), Lab Experiments are a Major Source of Knowledge in the Social Sciences, in: *Science* 326, 535–538

Fang, H./A. Moro (2011), Theories of Statistical Discrimination and Affirmative Action: A Survey, in: Benhabib, J./M. O. Jackson/A. Bisin (eds.), *Handbook of Social Economics*, Vol. 1A, North-Holland, 133–200

Farber, H. S./R. Gibbons (1996), Learning and Wage Dynamics, in: *Quarterly Journal of Economics* 111, 1007–1047

Farmer A./D. Terrell (1996), Discrimination, Bayesian Updating of Employer Beliefs and Human Capital Accumulation, in: *Economic Inquiry* 34, 204–219

Franzen, A./S. Pointner (2013), The External Validity of Giving in the Dictator Game: A Field Experiment Using the Misdirected Letter Technique, in: *Experimental Economics* 16, 155–169

Gaertner, S./L. Bickman (1971), Effects of Race on the Elicitation of Helping Behavior: The Wrong Number Technique, in: *Journal of Personality and Social Psychology* 20, 218–222

Gambetta, D. (2009), *Codes of the Underworld: How Criminals Communicate*, Princeton

Gerber, A. S./D. P. Green (2012), *Field Experiments: Design, Analysis, and Interpretation*, New York

Goffman, E. (1963), *Stigma: Notes on the Management of Spoiled Identity*, Englewood Cliffs

Goldberg, M. S. (1982), Discrimination, Nespotism, and Longrun Wage Differentials, in: *Quarterly Journal of Economic* 97, 307–319

Golder, S. A./M. W. Macy (2014), Digital Footprints: Opportunities and Challenges for Online Social Research, in: *Annual Review of Sociology* 40, 6–24

Guryan, J./K. Charles (2013), Taste-based or Statistical Discrimination: The Economics of Discrimination Returns to Its Roots, in: *Economic Journal* 123, F417–F432

Hainmueller, J./D. Hangartner/T. Yamamoto (2015), Validating Vignette and Conjoint Survey Experiments against Real-world Behavior, in: *Proceedings of the National Academy of Sciences of the United States of America* 112, 2395–2400

Heckman, J. J. (1998), Detecting Discrimination, in: *Journal of Economic Perspectives* 12, 101–116

— (2005), The Scientific Model of Causality, in: *Sociological Methodology* 35, 1–97

—/P. Siegelman (1993), The Urban Institute Audit Studies: Their Methods and Findings, in: Fix, M./R. Struyk (eds.), *Clear and Convincing Evidence: Measurement of Discrimination in America*, Washington, 187–258

Hedegaard, M./J.-R. Tyran (2014), The Price of Prejudice, in: *Discussion Paper No. 14-09, Department of Economics*, University of Copenhagen

Hedström, P. (2005), *Dissecting the Social. On the Principles of Analytical Sociology*, Cambridge

—/R. Swedberg (1998), Social Mechanisms: An Introductory Essay, in: Hedström, P./R. Swedberg (eds.), *Social Mechanisms: An Analytical Approach to Social Theory*, Cambridge, 1-31

—/P. Ylikoski (2010), Causal Mechanisms in the Social Sciences, in: *Annual Review of Sociology* 36, 49–67

Imai, K./L. Keele/D. Tingley/T. Yamamoto (2011), Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies, in: *American Political Science Review* 105, 765–789

—/D. Tingley/T. Yamamoto (2013), Experimental Designs for Identifying Causal Mechanisms, in: *Journal of the Royal Statistical Society A* 176, 5–51

Imbens, G. W./Rubin, D. B. (2015), *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, New York

Iversen, T./F. Rosenbluth (2010), *Women, Work, and Politics: The Political Economy of Gender Inequality*, Yale

Jackson, M./D. R. Cox (2013), The Principles of Experimental Design and Their Application in Sociology, in: *Annual Review of Sociology* 39, 27–49

Jasso, G./M. Webster Jr. (1999), Assessing the Gender Gap in Just Earnings and its Underlying Mechanisms, in: *Social Psychology Quarterly* 62, 367–380

Jowell, R./P. Prescott-Clarke (1970), Racial Discrimination and White-collar Workers in Britain, in: *Race and Class* 11, 397–417

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

Kalter, F. (2006), Auf der Suche nach einer Erklärung für die spezifischen Arbeitsmarktnachteile von Jugendlichen türkischer Herkunft, in: *Zeitschrift für Soziologie* 35, 144–160

Kalter, F./C. Kroneberg (2014), Between Mechanism Talk and Mechanism Cult: New Emphases in Explanatory Sociology and Empirical Research, in: Friedrichs, J./A. Nonnenmacher (eds.), *Kölner Zeitschrift für Soziologie und Sozialpsychologie, Special Issue* 54: Social Contexts and Social Mechanisms, 91–115

Keizer, K./S. Lindenberg/L. Steg (2008), The Spreading of Disorder, in: *Science* 322, 1681–1685

Kramer, A. D. I./J. E. Guillory/J. T. Hancock (2014), Experimental Evidence of Massive-scale Emotional Contagion through Social Networks, in: *Proceedings of the National Academy of Sciences of the United States of America* 111, 8788–8790

Krieger, N. (2014), Discrimination and Health Inequities, in: Berkman, L. F./I. Kawachi/M. Glymour (eds.), *Social Epidemiology*, 2nd edition, New York, 63–125

Krysan, M. (2000), Prejudice, Politics, and Public Opinion: Understanding the Sources of Racial Policy Attitudes, in: *Annual Review of Sociology* 26, 135–168

Lang, K./J.-Y. K. Lehmann (2012), Racial Discrimination in the Labor Market: Theory and Empirics, in: *Journal of Economic Literature* 50, 959–1006

LaPiere, R. T. (1934), Attitudes vs. Actions, in: *Social Forces* 13, 230–237

Levitt, S. D./J. A. List (2007), Viewpoint: On the Generalizability of Lab Behaviour to the Field, in: *Canadian Journal of Economics* 40, 347–370

—/— (2008), Field Experiments in Economics: The Past, the Present, and the Future, in: *European Economic Review* 53, 1–18

Lieberson, S. (1985), *Making It Count: The Improvement of Social Research and Theory*, Berkeley

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

Ludwig, J./J. R. Kling/S. Mullainathan (2011), Mechanism Experiments and Policy Evaluations, in: *Journal of Economic Perspectives* 25, 17–38

Merton, R. K. (1948), The Self-Fulfilling Prophecy, in: *Antioch Review* 8, 193–210

Midtbøen, A. H./J. Rogstad (2012), Discrimination: Methodological Controversies and Sociological Perspectives on Future Research, in: *Nordic Journal of Migration Research* 2, 203–212

Milgram, S./L. Mann/S. Harter (1965), The Lost Letter-Technique: A Tool of Social Research, in: *Public Opinion Quarterly* 29, 437–438

Mincer, J. A. (1974), *Schooling, Experience, and Earnings*, New York

Morgan, S. L./C. Winship (2015), *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, 2nd edition, Cambridge

Morton, R. B./K. Williams (2010), *Experimental Political Science and the Study of Causality. From Nature to the Lab*, Cambridge

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

Neyman, J./J. K. Iwaszkiewicz/S. Kolodziejczyk (1935), Statistical Problems in Agricultural Experimentation, in: *Supplement to the Journal of the Royal Statistical Society* 2, 107–180

Oaxaca, R. (1973), Male-female Wage Differentials in Urban Labor Markets, in: *International Economic Review* 14, 693–709

Pager, D./L. Quillian (2005), Walking the Talk? What Employers Say Versus What They Do, in: *American Sociological Review* 70, 355–380

—/D. Karafin (2009), Bayesian Bigot? Statistical Discrimination, Stereotypes, and Employer Decision Making, in: *Annals of the American Academy of Political and Social Sciences* 621, 70–93

—/H. Shepherd (2008), The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets, in: *Annual Review of Sociology* 34, 181–209

—/B. Western/B. Bonikowski (2009), Discrimination in a Low-Wage Labor Market: A Field Experiment, in: *American Sociological Review* 74, 777–799

Paradies, Y. (2006), A Systematic Review of Empirical Research on Self-Reported Racism and Health, in: *International Journal of Epidemiology* 35, 888–901

Petersen, T. (2009), Opportunities, in: Hedström, P./P. Bearman (eds.), *The Oxford Handbook of Analytical Sociology*, Oxford, 115–139

Phelps, E. (1972), The Statistical Theory of Racism and Sexism, in: *American Economic Review* 62, 659–661

Posner, R. (1995), *Aging and Old Age*, Chicago

Quillian, L. (2006), New Approaches to Understanding Racial Prejudice and Discrimination, in: *Annual Review of Sociology* 31, 299–328

Reskin, B. (2003), Including Mechanisms in our Models of Ascriptive Inequality, in: *American Sociological Review* 68, 1–21

Riach, P. A./J. Rich (2002), Field Experiments of Discrimination in the Market Place, in: *Economic Journal* 112, F480–F518

—/— (2004), Deceptive Field Experiments of Discrimination: Are They Ethical?, in: *Kyklos* 57, 457–470

Ridgeway, C. L. (1991), The Social Construction of Status Value: Gender and Other Nominal Characteristics, in: *Social Forces* 70, 367–380

Rubin, D. B. (1974), Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies, in: *Journal of Educational Psychology* 66, 688–701

— (1980), Randomization Analysis of Experimental Data: The Fisher Randomization Test, in: *Journal of the American Statistical Association* 75, 591–593

— (2008), For Objective Causal Inference, Design Trumps Analysis, in: *Annals of Applied Statistics* 2, 808–840

Schiermer, B. (2014), Late-modern Hipsters: New Tendencies in Popular Culture, in: *Acta Sociologica* 57, 167–181

Seibert, H./H. Solga (2005), Gleiche Chancen dank einer abgeschlossenen Ausbildung? Zum Signalwert von Ausbildungsabschlüssen bei ausländischen und deutschen jungen Erwachsenen, in: *Zeitschrift für Soziologie* 34, 364–382

Shadish, W. R./T. D. Cook/D. T. Campbell (2001), *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Boston

Sherif, M. (1958), Superordinate Goals in the Reduction of Intergroup Conflict, in: *American Journal of Sociology* 63, 349–356

Sobel, M. E. (2005), Discussion: 'The Scientific Model of Causality', in: *Sociological Methodology* 35, 99–133

Takács, K./F. Squazzoni/G. Bravo/M. Castellani (2014), Employer Networks, Priming, and Discrimination in Hiring: An Experiment, in: Manzo, G. (ed.), *Analytical Sociology: Norms, Actions, and Networks*, New York, 373–396

Tajfel, H. (1978), *Differentiation between Social Groups: Studies in the Social Psychology of Intergroup Relations*, London

—/J. C. Turner (1986), The Social Identity Theory of Intergroup Behavior, in: Worchel, S./W. G. Austin (eds.), *Psychology of Intergroup Relations*, Chicago, 7–24

Thomas, W. I./D. S. Thomas (1928), *The Child in America: Behavior Problems and Programs*, New York

Trautmann, S. T./G. van de Kuilen (2014), Belief Elicitation: A Horse Race among Truth Serums, in: *Economic Journal*, DOI: 10.1111/ecoj.12160

Williams, W. M./S. J. Ceci (2015), National Hiring Experiments Reveal 2:1 Faculty Preference for Women on STEM Tenure Track, in: *Proceedings of the National Academy of Sciences of the United States of America* 112, 5360–5365

Wilson, E. O. (2000), *Sociobiology: The New Synthesis*, Cambridge

Wolbring T./M. Keuschnigg (2015), Feldexperimente in den Sozialwissenschaften: Grundlagen, Herausforderungen, Beispiele, in: Keuschnigg, M./T. Wolbring (eds.), *Experimente in den Sozialwissenschaften*, Baden-Baden, 219–245

Wooldridge, J. (2004), *Estimating Average Partial Effects under Conditional Moment Independence Assumptions*, Cemmap Working Paper No. CWP03/04

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